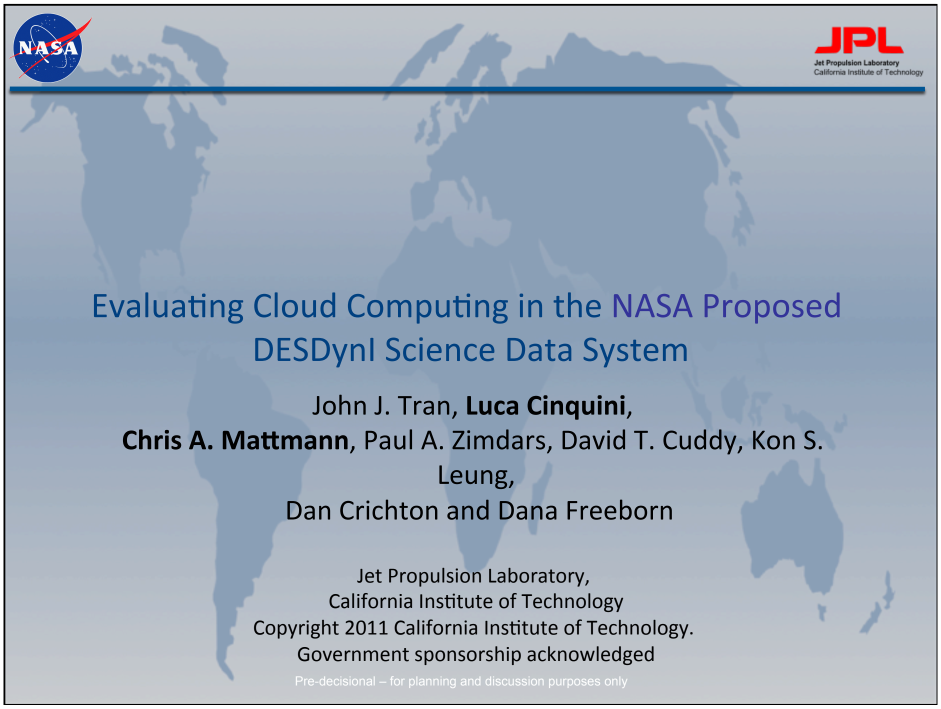


Evaluating Cloud Computing in the NASA Proposed DESDynI Science Data System

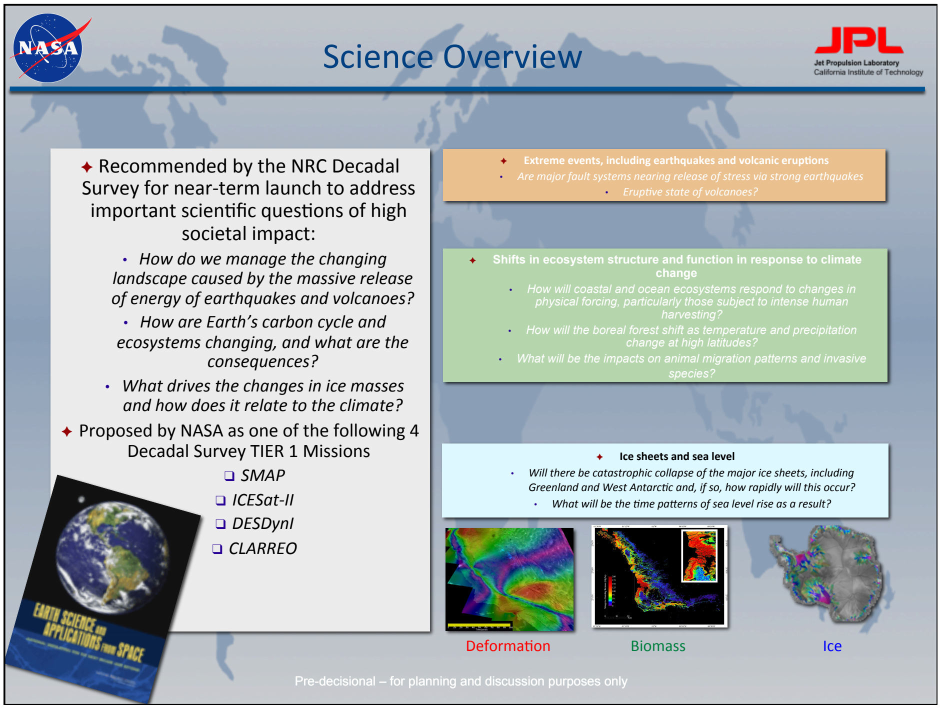
John J. Tran, **Luca Cinquini**,
Chris A. Mattmann, Paul A. Zimdars, David T. Cuddy, Kon S. Leung,
Dan Crichton and Dana Freeborn

Jet Propulsion Laboratory,
California Institute of Technology

Government sponsorship acknowledged

Agenda

Science Overview
Driving Requirements
Key Mission Interfaces
Architectural Study Goals
Results
Future Directions



Science Overview

✦ Recommended by the NRC Decadal Survey for near-term launch to address important scientific questions of high societal impact:

- *How do we manage the changing landscape caused by the massive release of energy of earthquakes and volcanoes?*
- *How are Earth's carbon cycle and ecosystems changing, and what are the consequences?*
- *What drives the changes in ice masses and how does it relate to the climate?*

✦ Proposed by NASA as one of the following 4 Decadal Survey TIER 1 Missions

❑ *SMAP*
❑ *ICESat-II*
❑ *DESDynI*
❑ *CLARREO*

✦ **Extreme events, including earthquakes and volcanic eruptions**

- *Are major fault systems nearing release of stress via strong earthquakes*
- *Eruptive state of volcanoes?*

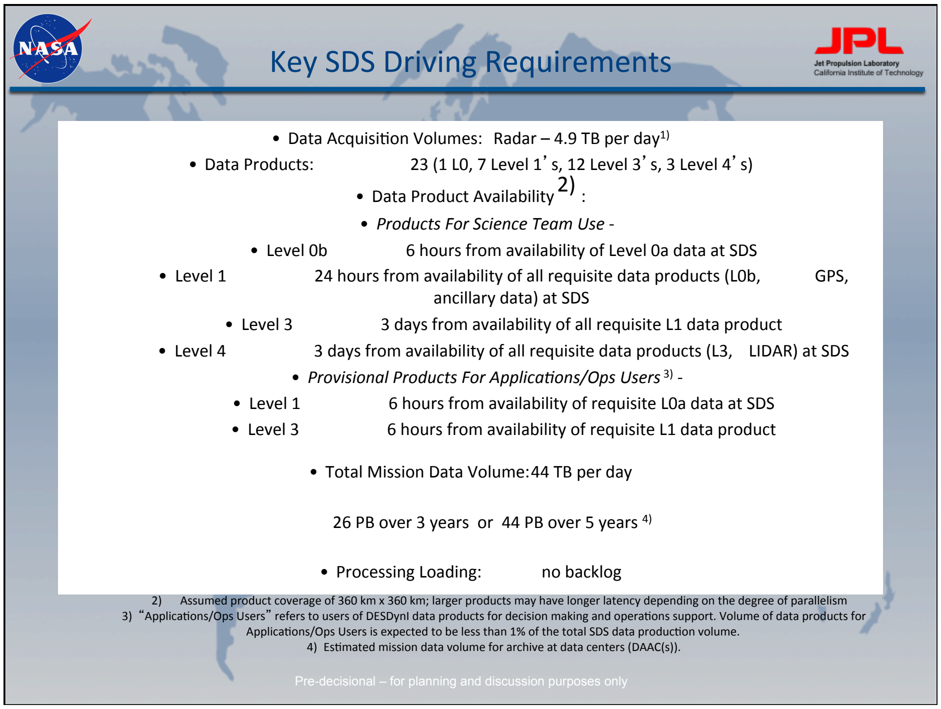
✦ **Shifts in ecosystem structure and function in response to climate change**

- *How will coastal and ocean ecosystems respond to changes in physical forcing, particularly those subject to intense human harvesting?*
- *How will the boreal forest shift as temperature and precipitation change at high latitudes?*
- *What will be the impacts on animal migration patterns and invasive species?*

✦ **Ice sheets and sea level**

- *Will there be catastrophic collapse of the major ice sheets, including Greenland and West Antarctic and, if so, how rapidly will this occur?*
- *What will be the time patterns of sea level rise as a result?*

Deformation Biomass Ice



Key SDS Driving Requirements

- Data Acquisition Volumes: Radar – 4.9 TB per day[1)]
- Data Products: 23 (1 L0, 7 Level 1's, 12 Level 3's, 3 Level 4's)
- Data Product Availability [2)] :
- *Products For Science Team Use -*
- Level 0b 6 hours from availability of Level 0a data at SDS
- Level 1 24 hours from availability of all requisite data products (L0b, GPS, ancillary data) at SDS
- Level 3 3 days from availability of all requisite L1 data product
- Level 4 3 days from availability of all requisite data products (L3, LIDAR) at SDS
- *Provisional Products For Applications/Ops Users* [3)] -
- Level 1 6 hours from availability of requisite L0a data at SDS
- Level 3 6 hours from availability of requisite L1 data product
- Total Mission Data Volume: 44 TB per day

26 PB over 3 years or 44 PB over 5 years [4)]

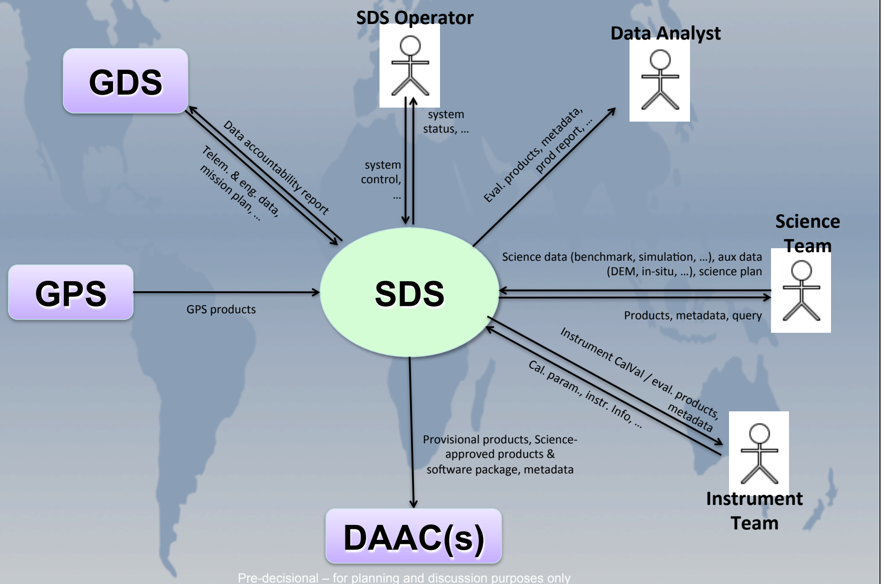
- Processing Loading: no backlog

[2)] Assumed product coverage of 360 km x 360 km; larger products may have longer latency depending on the degree of parallelism

[3)] "Applications/Ops Users" refers to users of DESDynI data products for decision making and operations support. Volume of data products for Applications/Ops Users is expected to be less than 1% of the total SDS data production volume.

[4)] Estimated mission data volume for archive at data centers (DAAC(s)).

SDS Key Interfaces


SDS Operator
Data Analyst
GDS
system status, …
system control, …
Data accountability report
Telem. & eng. data, mission plan, …
Eval. products, metadata, prod report, …
Science Team
GPS
GPS products
SDS
Science data (benchmark, simulation, …), aux data (DEM, in-situ, …), science plan
Products, metadata, query
Instrument Cal/val / eval. products, metadata
Cal. param., instr. Info, …
Provisional products, Science-approved products & software package, metadata
Instrument Team
DAAC(s)

Pre-decisional – for planning and discussion purposes only

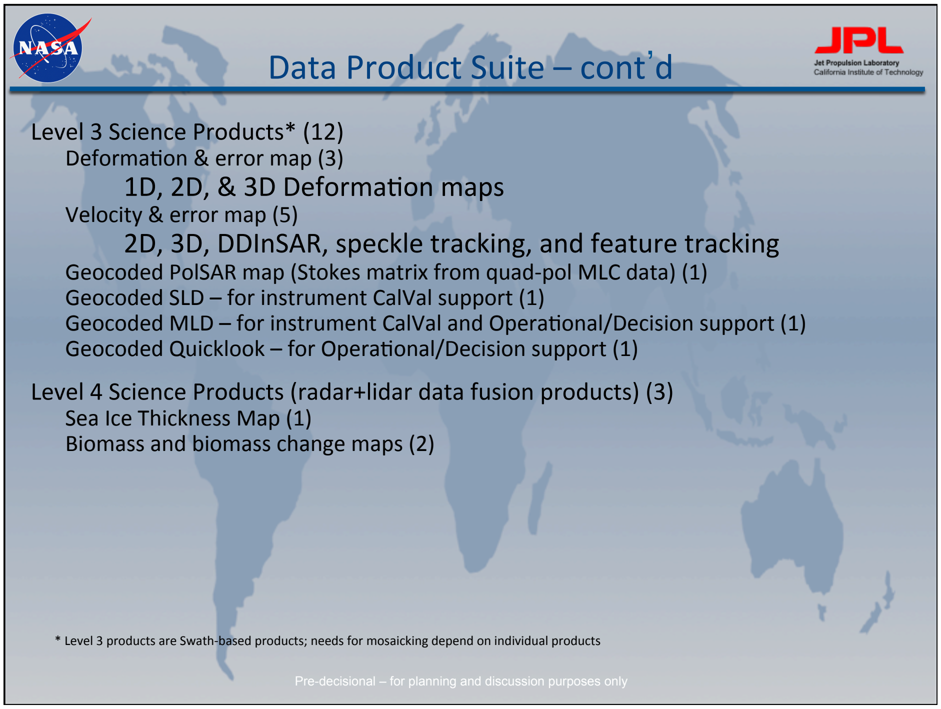
Data Product Suite

Level 0 Radar Instrument Products (1)
- Level 0b – Level 0a formatted for processing/archiving

Level 1 Radar Instrument Products (7)
- SLC – single-look complex (I/Q)
- SLD – detected SLC (amplitude)
- Quicklook product – uncalibrated detected image from L0b
- MLC – multi-look complex created from SLC by 4:1 averaging (2:1 in each dimension)
- MLD – detected MLC (amplitude)*
- Interferogram (I/Q; 4:1 averaging from full res interferogram)*
- Coherence (32-bit; the same dimension as the interferogram)*

*browse images may be generated as by-products

Pre-decisional – for planning and discussion purposes only



Data Product Suite – cont'd

Level 3 Science Products* (12)

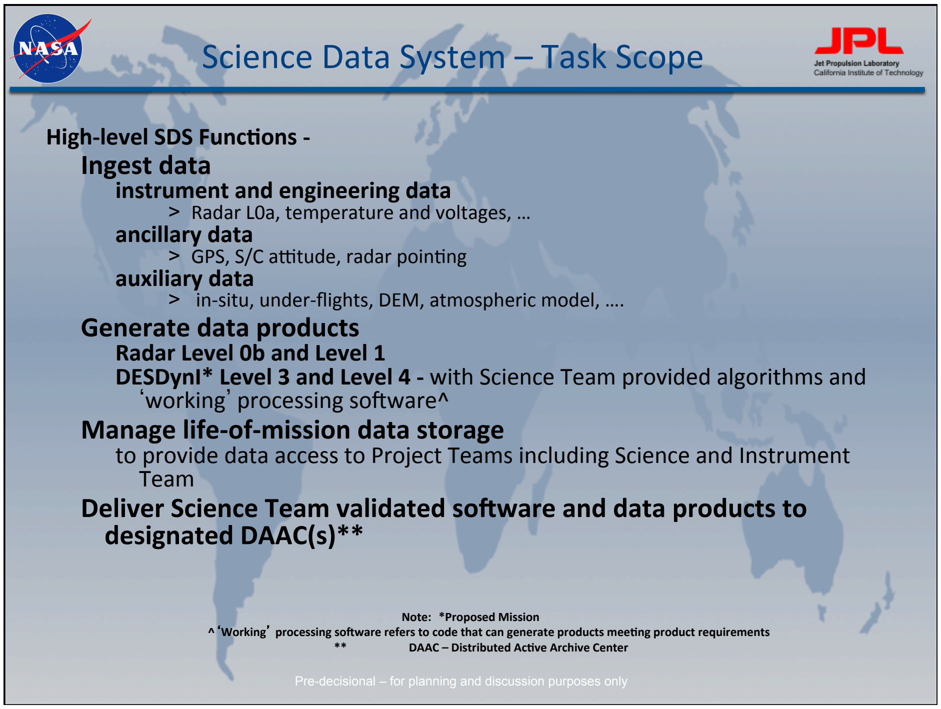
- Deformation & error map (3)
  - 1D, 2D, & 3D Deformation maps
- Velocity & error map (5)
  - 2D, 3D, DDInSAR, speckle tracking, and feature tracking
- Geocoded PolSAR map (Stokes matrix from quad-pol MLC data) (1)
- Geocoded SLD – for instrument CalVal support (1)
- Geocoded MLD – for instrument CalVal and Operational/Decision support (1)
- Geocoded Quicklook – for Operational/Decision support (1)

Level 4 Science Products (radar+lidar data fusion products) (3)

- Sea Ice Thickness Map (1)
- Biomass and biomass change maps (2)

* Level 3 products are Swath-based products; needs for mosaicking depend on individual products

Pre-decisional – for planning and discussion purposes only



Science Data System – Task Scope

High-level SDS Functions -

- **Ingest data**
  - **instrument and engineering data**
    - > Radar L0a, temperature and voltages, …
  - **ancillary data**
    - > GPS, S/C attitude, radar pointing
  - **auxiliary data**
    - > in-situ, under-flights, DEM, atmospheric model, ….
- **Generate data products**
  - **Radar Level 0b and Level 1**
  - **DESDynI* Level 3 and Level 4 -** with Science Team provided algorithms and 'working' processing software^
- **Manage life-of-mission data storage**
  - to provide data access to Project Teams including Science and Instrument Team
- **Deliver Science Team validated software and data products to designated DAAC(s)****

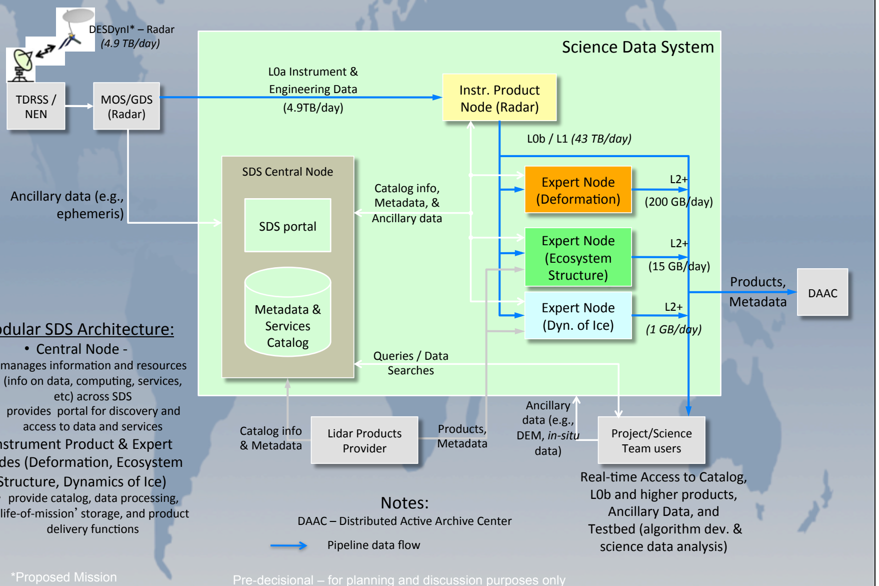
Note: *Proposed Mission

^ 'Working' processing software refers to code that can generate products meeting product requirements

**** DAAC – Distributed Active Archive Center**

Pre-decisional – for planning and discussion purposes only

SDS Functional Diagram


DESDynI* – Radar
(4.9 TB/day)
TDRSS / NEN
MOS/GDS (Radar)
Science Data System
L0a Instrument & Engineering Data
(4.9TB/day)
Instr. Product Node (Radar)
L0b / L1 *(43 TB/day)*
SDS Central Node
SDS portal
Metadata & Services Catalog
Ancillary data (e.g., ephemeris)
Catalog info, Metadata, & Ancillary data
Expert Node (Deformation)
L2+
(200 GB/day)
Expert Node (Ecosystem Structure)
L2+
(15 GB/day)
Expert Node (Dyn. of Ice)
L2+
(1 GB/day)
Products, Metadata
DAAC
Queries / Data Searches
Catalog info & Metadata
Lidar Products Provider
Products, Metadata
Ancillary data (e.g., DEM, *in-situ* data)
Project/Science Team users

Real-time Access to Catalog, L0b and higher products, Ancillary Data, and Testbed (algorithm dev. & science data analysis)

Modular SDS Architecture:

- Central Node -
  - manages information and resources (info on data, computing, services, etc) across SDS
  - provides portal for discovery and access to data and services
- Instrument Product & Expert Nodes (Deformation, Ecosystem Structure, Dynamics of Ice)
  - provide catalog, data processing, ‘life-of-mission’ storage, and product delivery functions

Notes:

DAAC – Distributed Active Archive Center

→ Pipeline data flow

*Proposed Mission

Pre-decisional – for planning and discussion purposes only

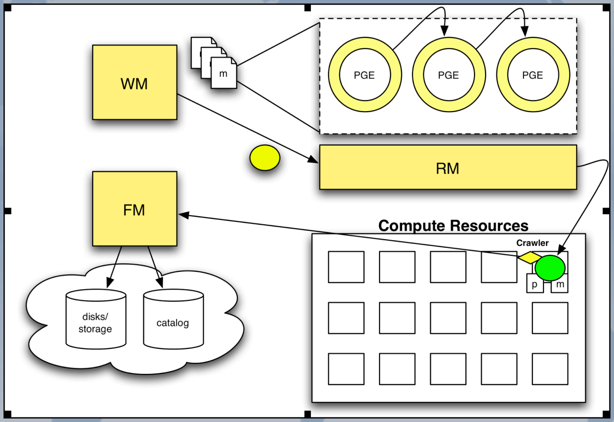
The Apache OODT framework

- Data System Infrastructure on which DESDynI*, SMAP*, NPP Sounder PEATE, OCO-2, ACOS, OCO were based
- Provides components for file management, workflow management, resource management
- Orchestrates science algorithms, cataloging and archiving their intermediate results, metadata, and scientific output
- Aka “Process Control System” or PCS

*Proposed Mission

Pre-decisional – for planning and discussion purposes only

Apache OODT in Action



Pre-decisional – for planning and discussion purposes only

Cloud Architectural Trade Study

Advance the SDS architectural design (**Goal G1)**

in addition to the overall SDS architecture, this item would also include the architectural aspects of the PCS/OODT with focus/emphasis on the specific nodes within the SDS

Investigate the trade spaces between central and distributed implementation (**Goal G2)**

better understand the parameters and landscapes of the trade spaces available

see if there exists an optimal approach that best suits DESDynI's proposed goals and objectives

Revisit the data generation, storage, backup, and distribution strategies (**Goal G3**)

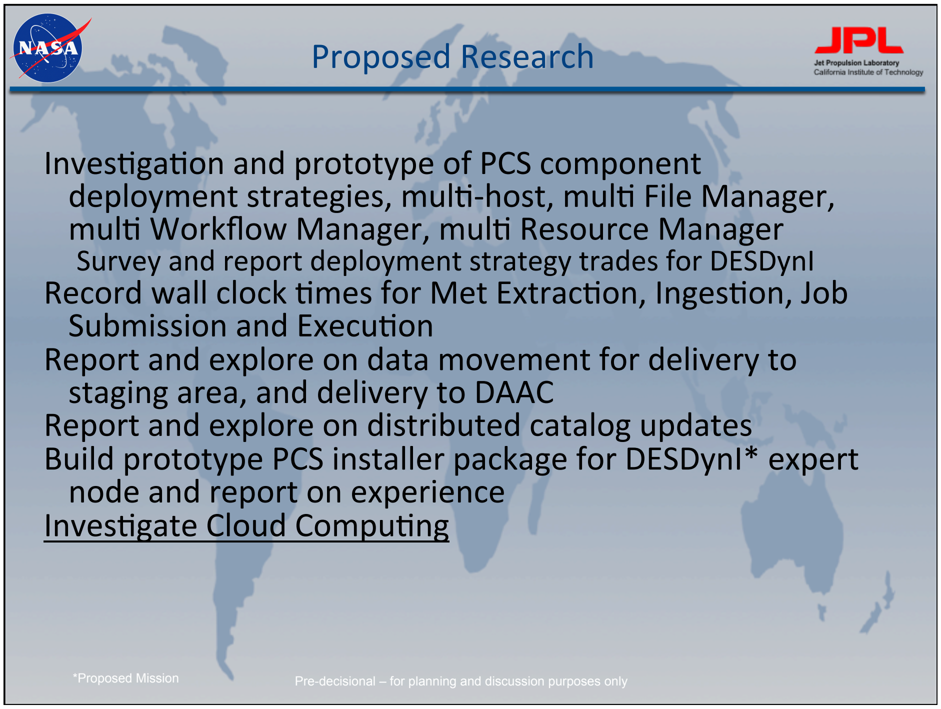
surveying the latest and projected trends and technologies in Computing & Data Storage

identifying and evaluating plausible strategies applicable to the DESDynI concept

in view of the varying time-series nature of some of the products, including trade(s) between product storage and product re-generation

recommending an optimal (most efficient and cost effective) data generation/storage/backup/distribution strategy that would best suit DESDynI's proposed goals and objectives

Pre-decisional – for planning and discussion purposes only

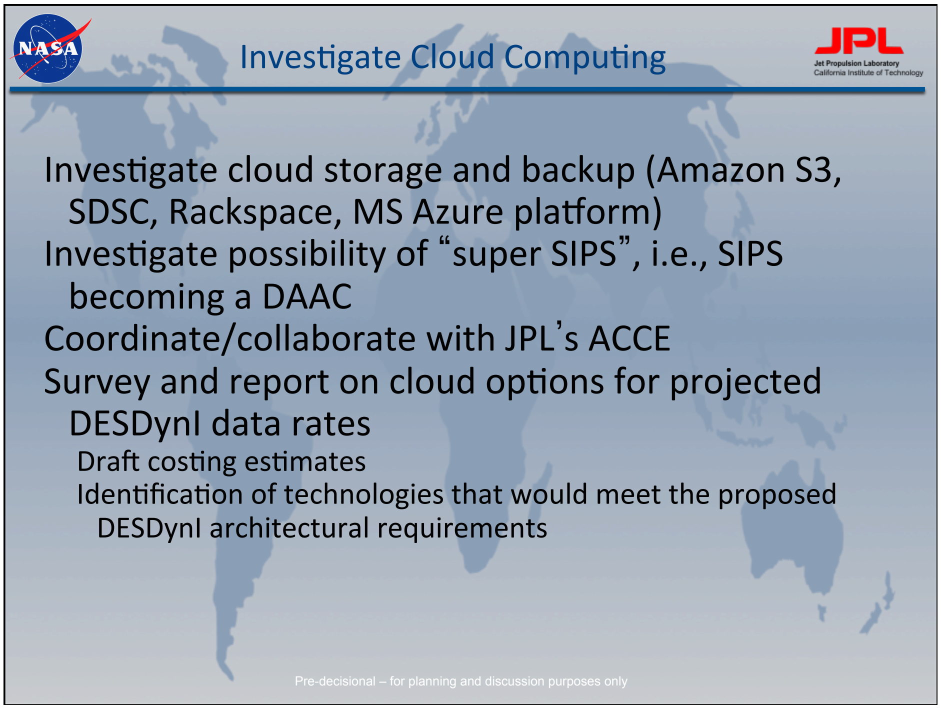


Proposed Research

- Investigation and prototype of PCS component deployment strategies, multi-host, multi File Manager, multi Workflow Manager, multi Resource Manager
  - Survey and report deployment strategy trades for DESDynI
- Record wall clock times for Met Extraction, Ingestion, Job Submission and Execution
- Report and explore on data movement for delivery to staging area, and delivery to DAAC
- Report and explore on distributed catalog updates
- Build prototype PCS installer package for DESDynI* expert node and report on experience
- <u>Investigate Cloud Computing</u>

*Proposed Mission

Pre-decisional – for planning and discussion purposes only



Investigate Cloud Computing

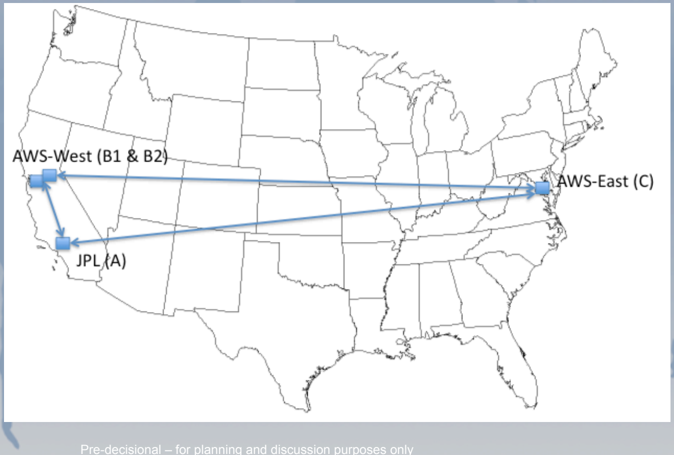
- Investigate cloud storage and backup (Amazon S3, SDSC, Rackspace, MS Azure platform)
- Investigate possibility of “super SIPS”, i.e., SIPS becoming a DAAC
- Coordinate/collaborate with JPL’s ACCE
- Survey and report on cloud options for projected DESDynI data rates
  - Draft costing estimates
  - Identification of technologies that would meet the proposed DESDynI architectural requirements

Pre-decisional – for planning and discussion purposes only

Decided on leveraging Amazon

Simple, JPL had close collaboration with Amazon provider

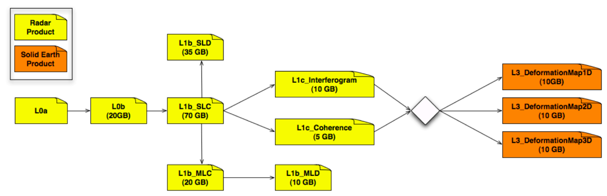
JPL internal VPN to AWS backbone



Pre-decisional – for planning and discussion purposes only

Wanted to benchmark real pipeline

Picked subset of overall projected DESDynI pipeline

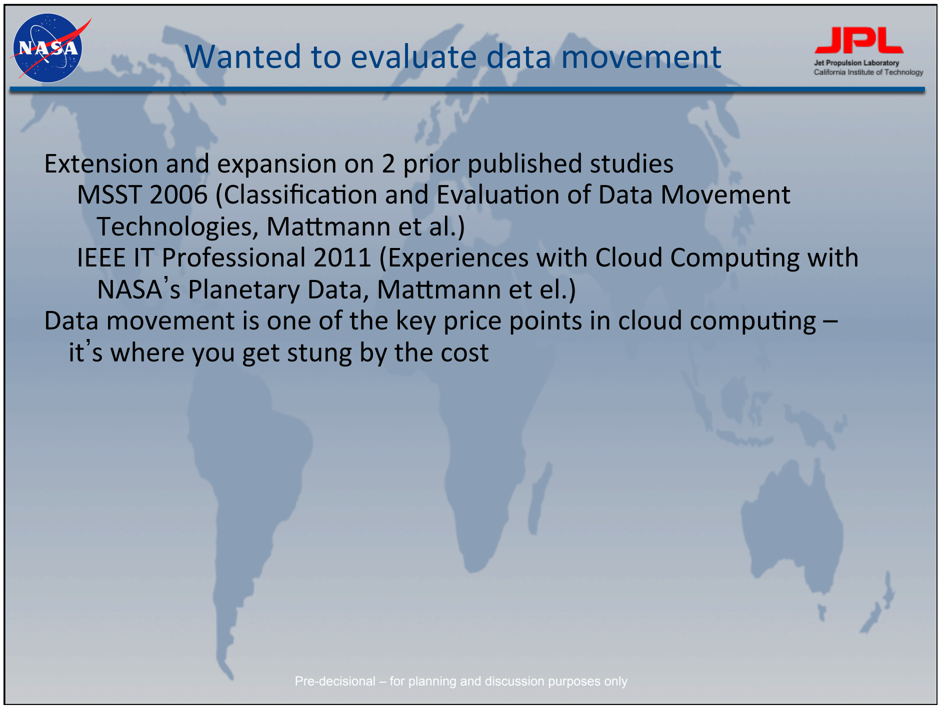


Representative of fan-in, fan-out complexity

Also representative of sizing

Data sizing and transfer one of the important aspects to benchmark

Pre-decisional – for planning and discussion purposes only



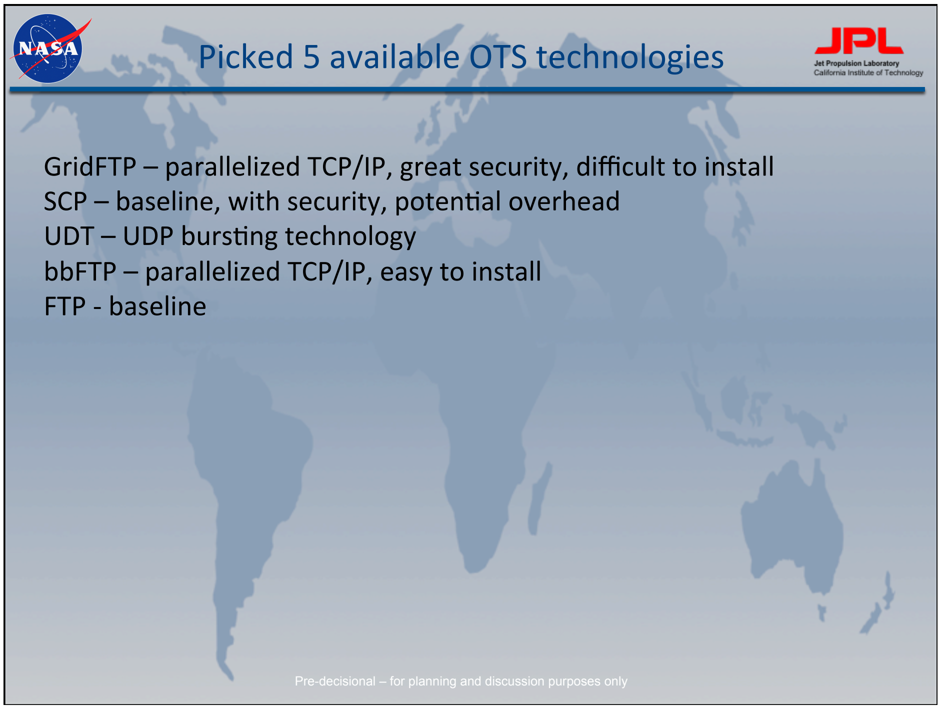
Wanted to evaluate data movement

Extension and expansion on 2 prior published studies

- MSST 2006 (Classification and Evaluation of Data Movement Technologies, Mattmann et al.)
- IEEE IT Professional 2011 (Experiences with Cloud Computing with NASA's Planetary Data, Mattmann et el.)

Data movement is one of the key price points in cloud computing – it's where you get stung by the cost

Pre-decisional – for planning and discussion purposes only



Picked 5 available OTS technologies

GridFTP – parallelized TCP/IP, great security, difficult to install

SCP – baseline, with security, potential overhead

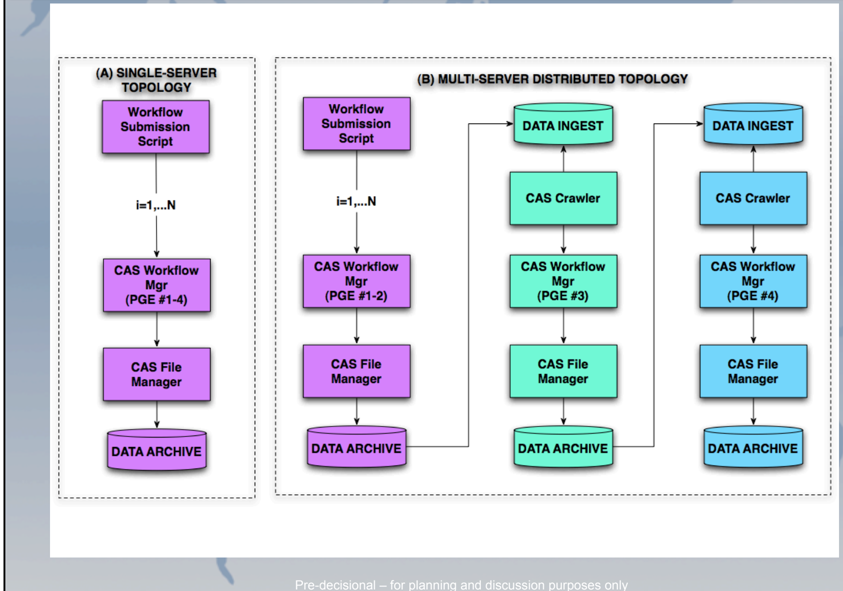
UDT – UDP bursting technology

bbFTP – parallelized TCP/IP, easy to install

FTP - baseline

Pre-decisional – for planning and discussion purposes only

Overall DESDynI topologies tested


(A) SINGLE-SERVER TOPOLOGY
Workflow Submission Script
i=1,...N
CAS Workflow Mgr (PGE #1-4)
CAS File Manager
DATA ARCHIVE
(B) MULTI-SERVER DISTRIBUTED TOPOLOGY
Workflow Submission Script
i=1,...N
CAS Workflow Mgr (PGE #1-2)
CAS File Manager
DATA ARCHIVE
DATA INGEST
CAS Crawler
CAS Workflow Mgr (PGE #3)
CAS File Manager
DATA ARCHIVE
DATA INGEST
CAS Crawler
CAS Workflow Mgr (PGE #4)
CAS File Manager
DATA ARCHIVE

Desdyni Data Processing Pipeline Studies

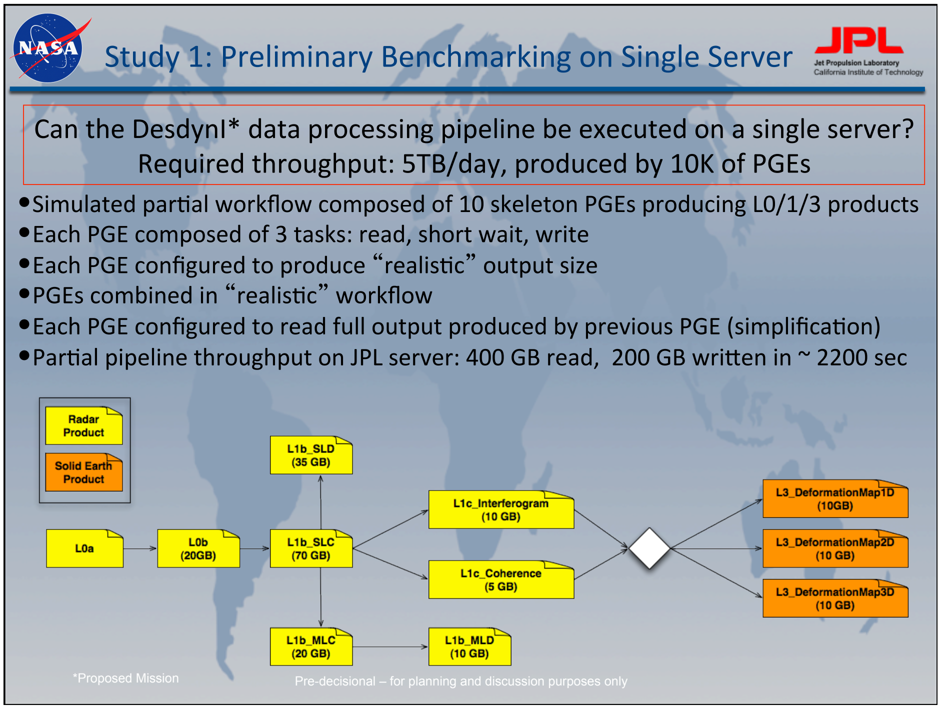
3 simulation studies were conducted to advance our knowledge of the optimal DesdynI* data system architecture:

- Study 1: <u>Preliminary benchmarking on single server</u>
  - Execute a ball-park estimate of the maximum possible pipeline throughput when executed onto a single server

- Study 2: <u>Compare available data transfer technologies</u>
  - Research and benchmark data transfer technologies in a cloud environment to assess wether data transfer is a limiting factor

- Study 3: <u>Analyze tradeoffs of centralized versus distributed architecture</u>
  - Deploy data processing pipeline in different topologies in a cloud environment, and compare overall execution time

*Proposed Mission



Study 1: Preliminary Benchmarking on Single Server

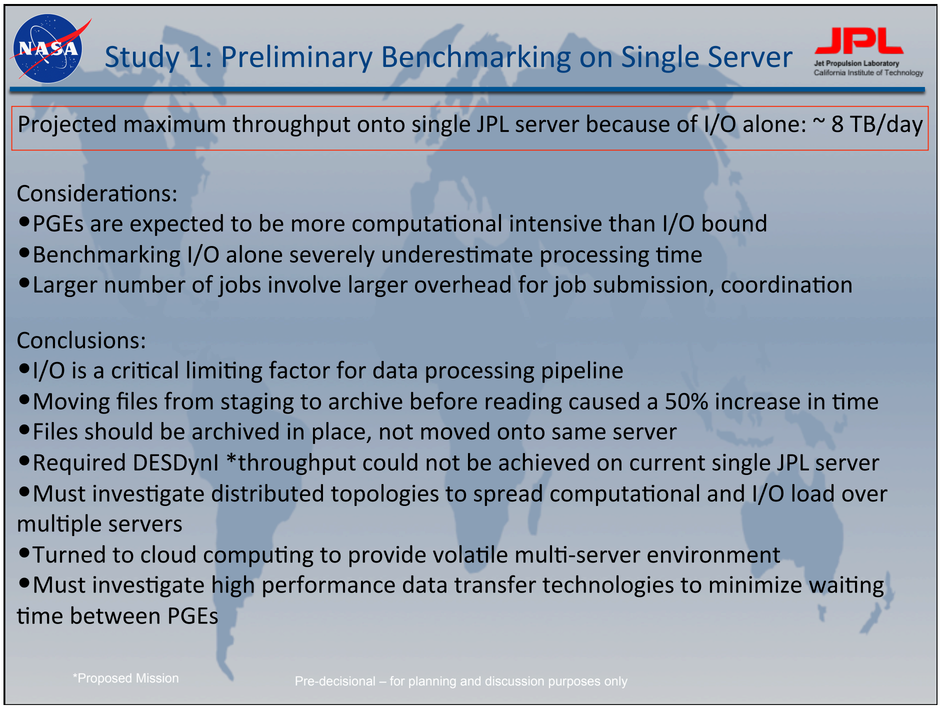
Can the DesdynI* data processing pipeline be executed on a single server?
Required throughput: 5TB/day, produced by 10K of PGEs

- Simulated partial workflow composed of 10 skeleton PGEs producing L0/1/3 products
- Each PGE composed of 3 tasks: read, short wait, write
- Each PGE configured to produce "realistic" output size
- PGEs combined in "realistic" workflow
- Each PGE configured to read full output produced by previous PGE (simplification)
- Partial pipeline throughput on JPL server: 400 GB read, 200 GB written in ~ 2200 sec

Radar Product
Solid Earth Product
L0a
L0b (20GB)
L1b_SLD (35 GB)
L1b_SLC (70 GB)
L1b_MLC (20 GB)
L1b_MLD (10 GB)
L1c_Interferogram (10 GB)
L1c_Coherence (5 GB)
L3_DeformationMap1D (10GB)
L3_DeformationMap2D (10 GB)
L3_DeformationMap3D (10 GB)

*Proposed Mission

Pre-decisional – for planning and discussion purposes only



Study 1: Preliminary Benchmarking on Single Server

Projected maximum throughput onto single JPL server because of I/O alone: ~ 8 TB/day

Considerations:

- PGEs are expected to be more computational intensive than I/O bound
- Benchmarking I/O alone severely underestimate processing time
- Larger number of jobs involve larger overhead for job submission, coordination

Conclusions:

- I/O is a critical limiting factor for data processing pipeline
- Moving files from staging to archive before reading caused a 50% increase in time
- Files should be archived in place, not moved onto same server
- Required DESDynI *throughput could not be achieved on current single JPL server
- Must investigate distributed topologies to spread computational and I/O load over multiple servers
- Turned to cloud computing to provide volatile multi-server environment
- Must investigate high performance data transfer technologies to minimize waiting time between PGEs

*Proposed Mission

Pre-decisional – for planning and discussion purposes only

Experimental Setup for Cloud Computing Studies

- Selected Amazon AWS as cloud computing vendor
  - JPL is connected by high speed internet to AWS West network
  - Could leverage experience of other JPL groups (such as ACCE) w/ AWS
- Integrated the full simulated DESDynI* software stack into one single package (OODT/CAS, multiple data transfer technologies and simulated science processing algorithms)
  - Used EC2 developer toolkit to create one master AMI (Amazon Machine Image)
  - High-end cloud server (type: m2.4xlarge, high I/O performance)
  - Memory: 68.4 GB
  - CPU: 26 EC2 Compute Nodes (corresponding to 8 virtual cores)
  - Internal disk: 1690 GB
  - Replicated AMI setup multiple times as needed

Lessons learned:

- AMI configuration must be executed from the command line, no GUI available
- AMI replication across Amazon zones (West to East) is non-trivial and not well documented

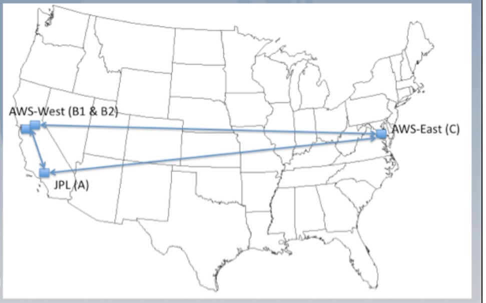
*Proposed Mission

Study 2: Compare Data Transfer Technologies

- Researched 5 open source technologies:
  - FTP: widely used, no strong security, baseline
  - scp: ubiquitous, built-in SSH security, potential encryption overhead
  - GridFTP: parallelized TCP/IP, strong security, complex installation & configuration
  - bbFTP: parallelized TCP/IP, easy installation (standalone client/server)
  - UDT: UDP (User Datagram Protocol) bursting technology
  - Transferred 1GB, 10GB NetCDF files (compressed) between JPL (A) and Amazon WS clusters
  - AWS-West B1 & B2, internal and external network
  - AWS-East (C), external network

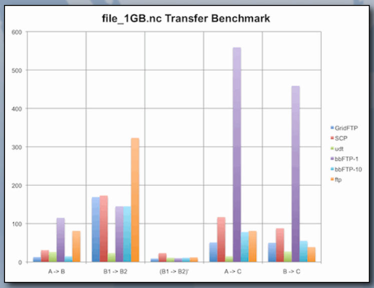
Previous Studies References:

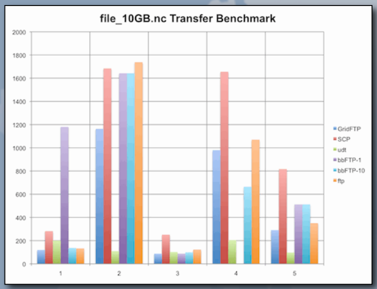
- MSST 2006 (“Classification and Evaluation of Data Movement Technologies”, Mattmann et al.)
- IEEE IT Professional 2011 (“Experiences with Cloud Computing with NASA’s Planetary Data”, Mattmann et el.)



Study 2: Compare Data Transfer Technologies

Jet Propulsion Laboratory
California Institute of Technology


file_1GB.nc Transfer Benchmark


file_10GB.nc Transfer Benchmark

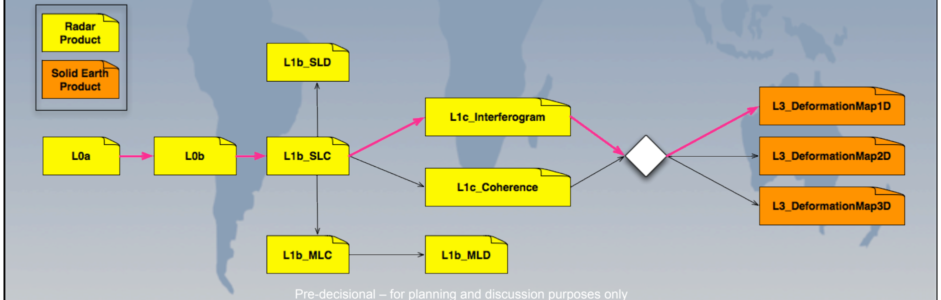
- Measured transfer rates between AWS nodes vary considerably, likely due to concurrent use of servers, network by other projects
- Using internal Amazon network consistently yields better performance than when using the public network (as determined by the node IP addresses)
- UDT and GridFTP offer best performance for both 1GB and 10GB size files
- GridFTP is recommended if security and data integrity are required
- UDT is recommended if easiness of installation is a priority
- Note: cloud providers charge considerably for data transfer (and data storage)

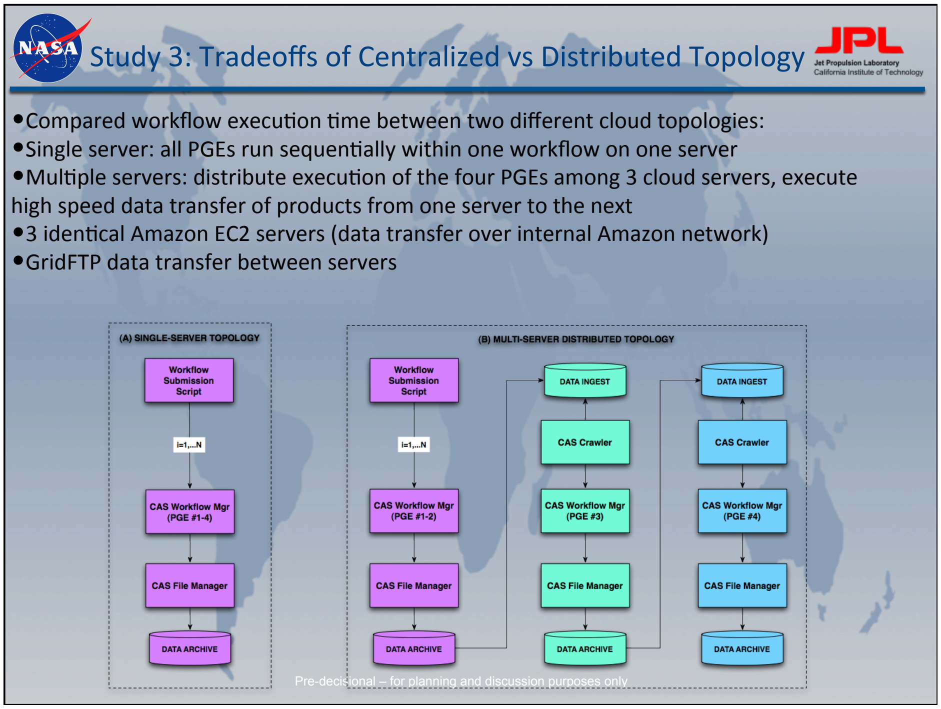
Study 3: Tradeoffs of Centralized vs Distributed Topology

Jet Propulsion Laboratory
California Institute of Technology

Goal: leverage cloud computing to evaluate topology tradeoffs for DESDynI architecture: single server versus multiple servers deployments

- Defined a data processing pipeline composed of 4 ROI-PAC PGEs
  - ROIPAC: open source software package for processing interferometry data
  - Successive generation of level 1, 2 and 3 data products
  - Output from one PGE used as input for next PGE
  - Considerably I/O and computing intensive
  - Real science processing tasks, representatives of projected DESDynI tasks



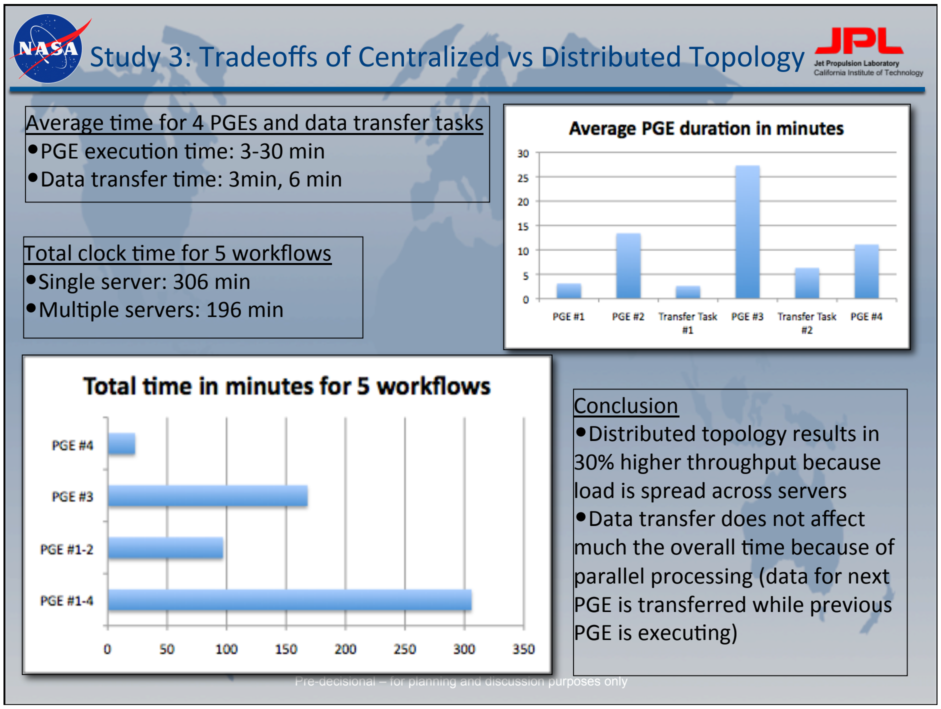


Study 3: Tradeoffs of Centralized vs Distributed Topology

- Compared workflow execution time between two different cloud topologies:
- Single server: all PGEs run sequentially within one workflow on one server
- Multiple servers: distribute execution of the four PGEs among 3 cloud servers, execute high speed data transfer of products from one server to the next
- 3 identical Amazon EC2 servers (data transfer over internal Amazon network)
- GridFTP data transfer between servers

(A) SINGLE-SERVER TOPOLOGY
Workflow Submission Script
i=1,...N
CAS Workflow Mgr (PGE #1-4)
CAS File Manager
DATA ARCHIVE
(B) MULTI-SERVER DISTRIBUTED TOPOLOGY
Workflow Submission Script
i=1,...N
CAS Workflow Mgr (PGE #1-2)
CAS File Manager
DATA ARCHIVE
DATA INGEST
CAS Crawler
CAS Workflow Mgr (PGE #3)
CAS File Manager
DATA ARCHIVE
DATA INGEST
CAS Crawler
CAS Workflow Mgr (PGE #4)
CAS File Manager
DATA ARCHIVE

Pre-decisional – for planning and discussion purposes only



Study 3: Tradeoffs of Centralized vs Distributed Topology

Average time for 4 PGEs and data transfer tasks
- PGE execution time: 3-30 min
- Data transfer time: 3min, 6 min

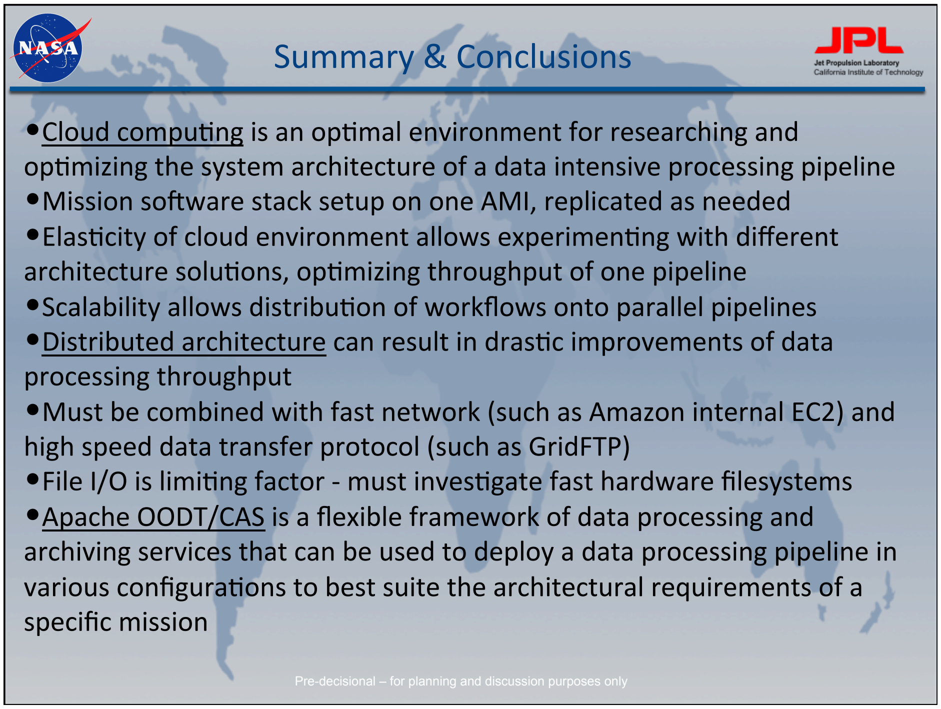
Total clock time for 5 workflows
- Single server: 306 min
- Multiple servers: 196 min

Average PGE duration in minutes

Total time in minutes for 5 workflows

Conclusion
- Distributed topology results in 30% higher throughput because load is spread across servers
- Data transfer does not affect much the overall time because of parallel processing (data for next PGE is transferred while previous PGE is executing)

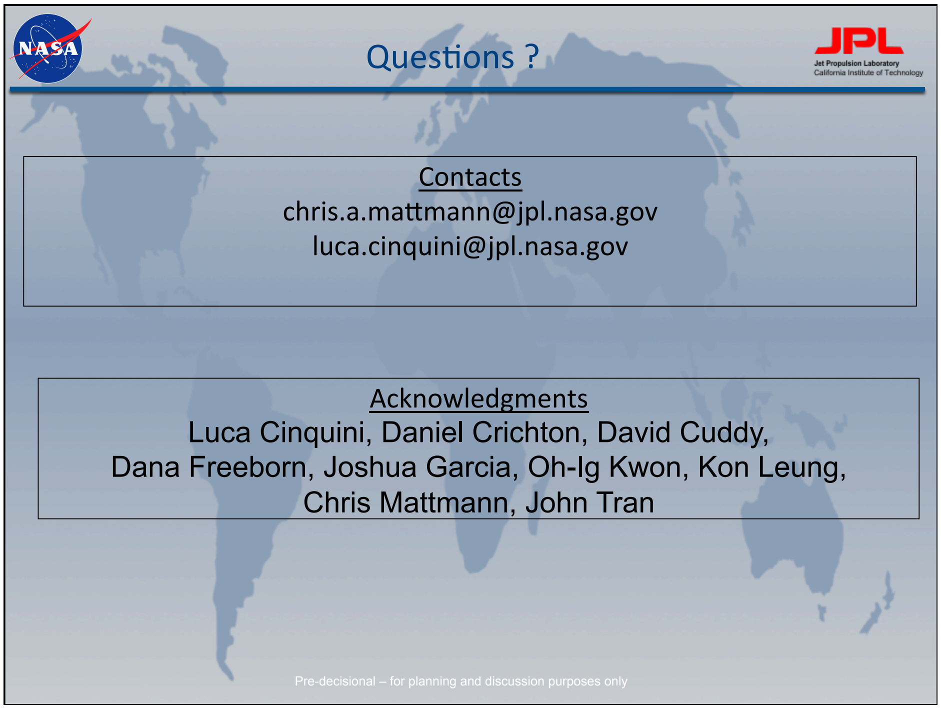
Pre-decisional – for planning and discussion purposes only



Summary & Conclusions

- Cloud computing is an optimal environment for researching and optimizing the system architecture of a data intensive processing pipeline
- Mission software stack setup on one AMI, replicated as needed
- Elasticity of cloud environment allows experimenting with different architecture solutions, optimizing throughput of one pipeline
- Scalability allows distribution of workflows onto parallel pipelines
- Distributed architecture can result in drastic improvements of data processing throughput
- Must be combined with fast network (such as Amazon internal EC2) and high speed data transfer protocol (such as GridFTP)
- File I/O is limiting factor - must investigate fast hardware filesystems
- Apache OODT/CAS is a flexible framework of data processing and archiving services that can be used to deploy a data processing pipeline in various configurations to best suite the architectural requirements of a specific mission

Pre-decisional – for planning and discussion purposes only



Questions ?

Contacts
chris.a.mattmann@jpl.nasa.gov
luca.cinquini@jpl.nasa.gov

Acknowledgments
Luca Cinquini, Daniel Crichton, David Cuddy,
Dana Freeborn, Joshua Garcia, Oh-Ig Kwon, Kon Leung,
Chris Mattmann, John Tran

Pre-decisional – for planning and discussion purposes only