

---

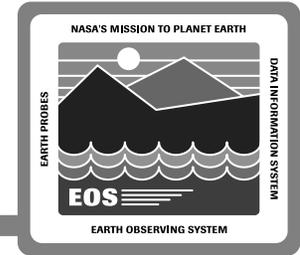
# Science Data Preprocessing CSCI

## Narayan Prasad

---

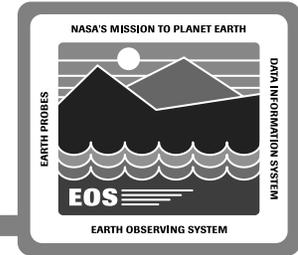
15 February 1995

# Data Preprocessing (DPREP)

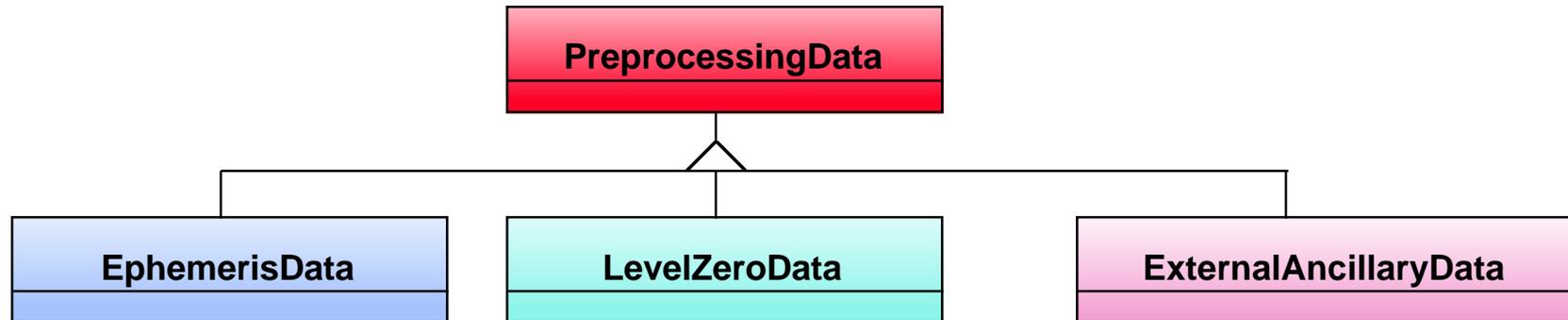


- **Scope of extent will change based on the resolution of some L3 requirements**
- **Collected Preprocessing functions into a single CSCI**
- **Results of our analysis show that we can actually handle the required functionality by distributing it across other CSCIs**
- **Design concept is to facilitate identification and distribution to appropriate subsystem with minimal impact on their design**
- **Data types that may need preprocessing**
  - **O/A data in spacecraft ancillary data for TRMM and EOS-AM L0**
  - **Flight Dynamics Facility (FDF)-generated ephemeris data for TRMM and EOS-AM**
  - **External ancillary data (e.g., NOAA)**

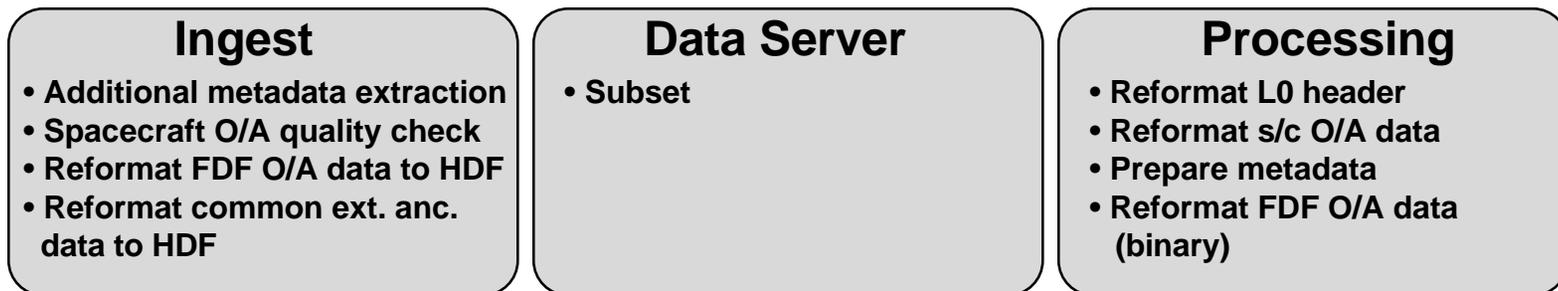
# Design Approach/Rationale



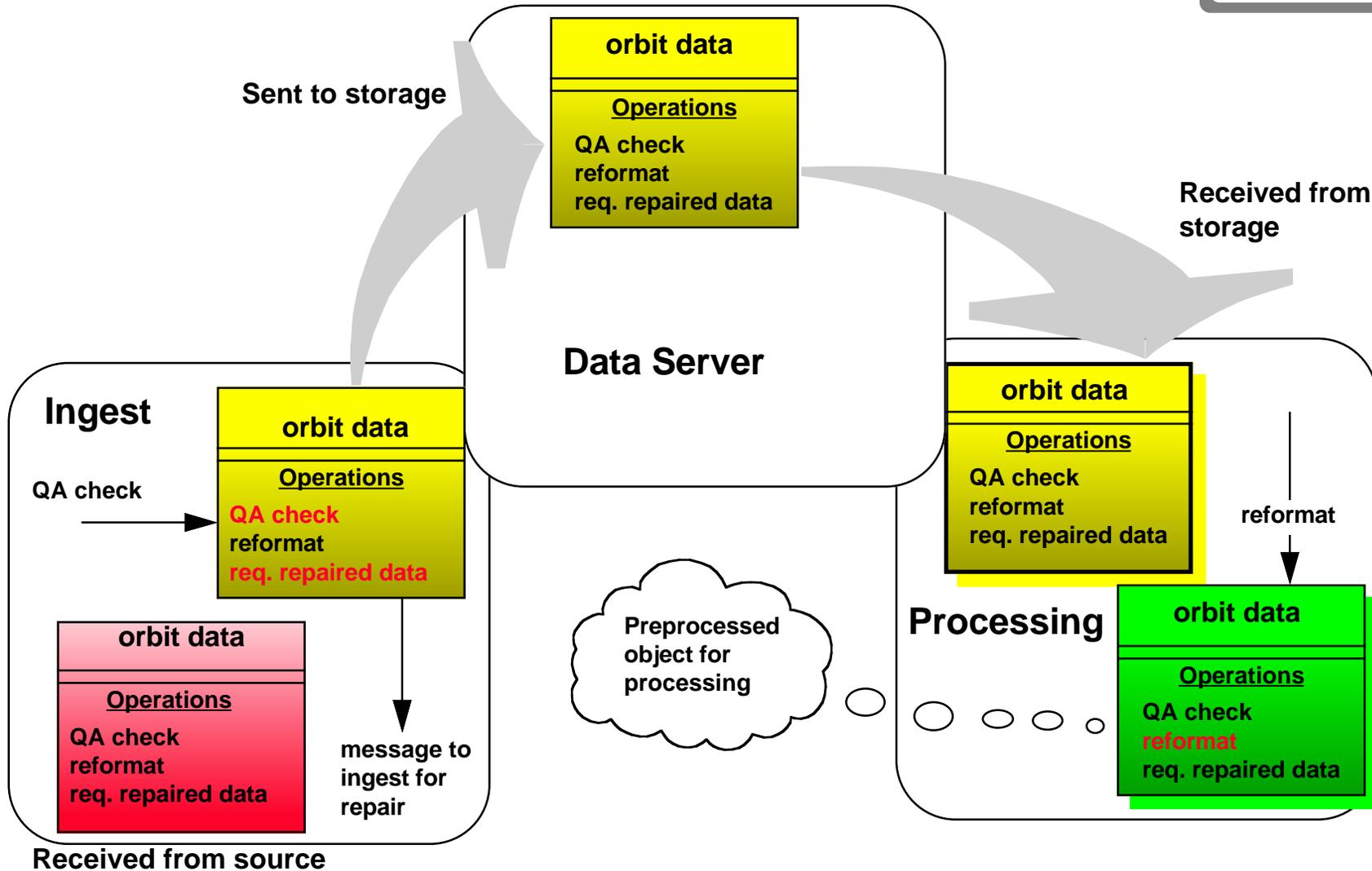
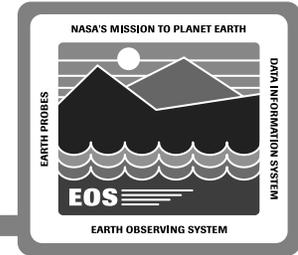
- Modeled as a class library with interfaces as software functions and arguments called by each subsystem
- Data types hierarchy



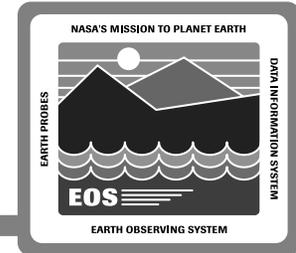
## Proposed distribution of Preprocessing functions across subsystems:



# Design Approach/Rationale (cont.)



# Road Map for Planning & Processing Presentation



## Overview

- Concept Drivers, Key Features
- Production Management Flow

## Software Model

## COTS/Prototypes

- Evaluation
- OTS and Software Reuse

## Phasing of Capabilities

## Scenarios

## Cross DAAC Scheduling/Planning

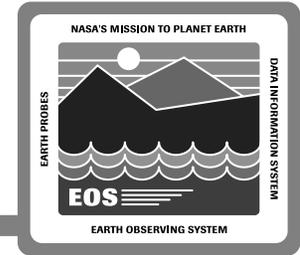
## Other Data Processing CIs

- AI&T Tools
- Science Data Preprocessing

## *Hardware*

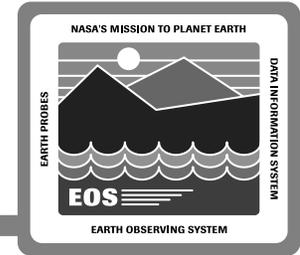
## Issues

# Overview



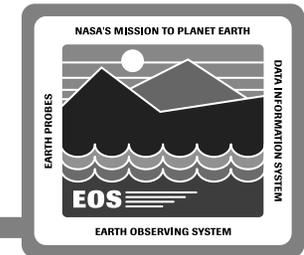
- **Purpose**
  - **Provide processing resources, system software, and COTS packages to support Science Processing, Algorithm Integration & Test, System Integration & Test, and Q/A**
  - **Provide Planning hardware to support Production Planning, Production Management, and Data Server/Ingest Access**
- **Focus**
  - **IR-1 and Release A**
  - **“Look Ahead” to Release B for Scalability and Evolvability**
  - **DAAC unique**

# Design Drivers



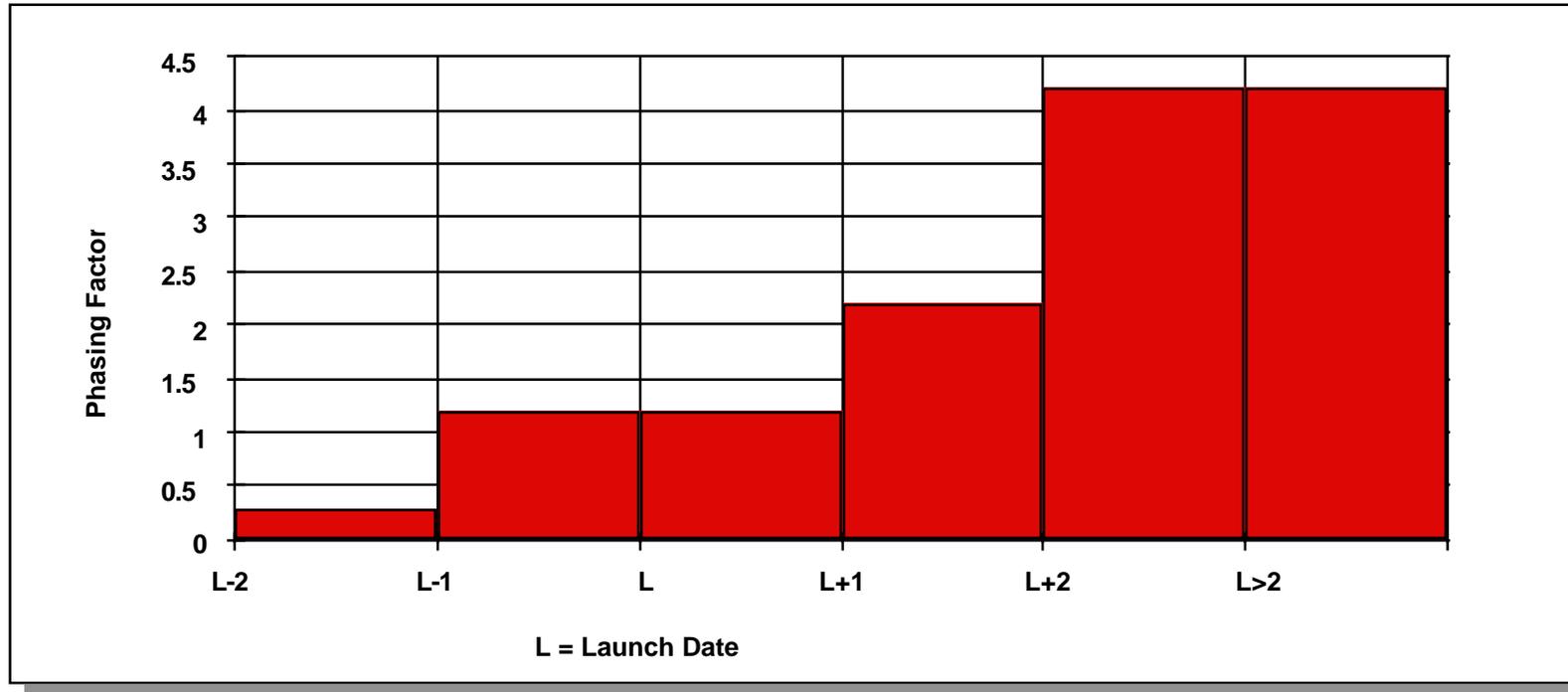
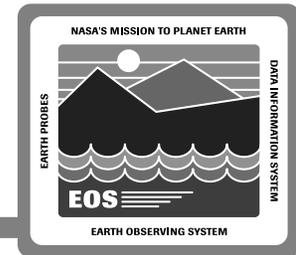
- **Derived from AHWGP data**
  - **processing, disk storage, I/O bandwidth, PGE activations**
- **Separation of AI&T environment from operations environment**
- **Provide support for the integration of science software**
- **Ease of algorithm development, integration and maintenance**
  - **learning curve, cost, skill mix**
  - **minimize software modification**
- **Scalability and evolvability to Release B and beyond**

# Design Rationale



- **Hardware recommendation**
  - analysis of AHWGP data
  - preliminary results from System Performance Model (Release A)
  - ESDIS Phasing Factors (pre-launch and post-launch factors)
  - ECS Science and Technology Lab Prototyping
  - trade studies (DID 211)
    1. Distributed and Parallel Processing
    2. Platform Families
    3. Production Topology
  - price/performance tradeoffs
- **Design supports**
  - phased procurement
  - heterogeneous architectures (uniprocessor, SMP, Workstation cluster, and MPP)
  - use of heritage science software
  - multivendor platform

# ESDIS Phasing Factors



## •0.3 X starting at L-2 years

- X is defined as at-launch processing estimate for pre-launch AI & T

## •1.2 X starting at L-1 years

- X is defined as at-launch processing estimate for pre-launch AI & T and systems I & T

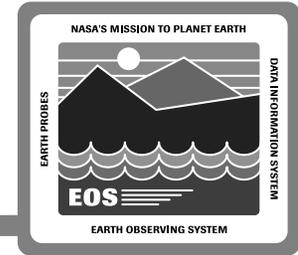
## •2.2 X starting at L+1 years

- X is defined as quarterly processing estimate for post-launch period

## •4.2 X starting starting at L+2 years

- X is defined as quarterly processing estimate for post-launch period

# Prototype Studies Highlights



## Prototype

### Science Software Execution Prototype

- Used science algorithms (e.g., Pathfinder AVHRR/Land, SSM/I, SeaWinds to study applicability of various processing alternatives (DCE, SMP, DMP/workstation cluster and MPP)

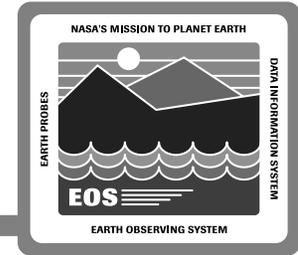
## Usefulness for ECS

- Provided inputs for AI&T
  - portability (32-64 bit issues)
- Revealed new processing technologies
  - architectures and software tools
- Provided lessons to select hardware
- Prototyped SSM/I parallel code for SMP will soon be provided to MSFC DAAC to study suitability for production

## Features

- Distributed computing of Pathfinder AVHRR/Land using OSF/DCE on geographically distributed workstation cluster
- Multiprocessing (using SMP, DMP/workstation cluster and MPP) of SSM/I and SeaWinds using automatic parallelization tools
- SDP Toolkit performance studies

# Trade Studies Highlights



## Trade studies

- **Distributed/Parallel Processing**
  - studies various processing alternatives for ECS science algorithms
- **Production Platform Families**
  - based on the Technical Baseline/ AHWGP data, scalability, cost, etc., recommend one or more Processing subsystem processor classes
- **Production Topology**
  - analyzes physical (not logical) processing topologies that can impact hardware requirements, overall performance, network capacity, throughput and staging storage

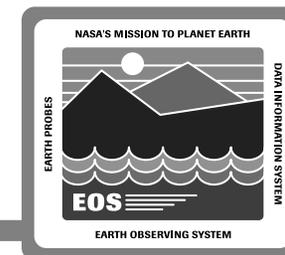
## Usefulness for ECS

- **Distributed/Parallel Computing**
  - provides up-to-date information on processing technologies
- **Production Platform Families**
  - provides basis and rationale for hardware selection for IR1 and Release A
- **Production Topology Results**
  - will be used for Release B hardware selection

## Features

- **Production Topology Trade**
  - recommendations for hardware selection based on cluster optimization alternatives for Release B and beyond

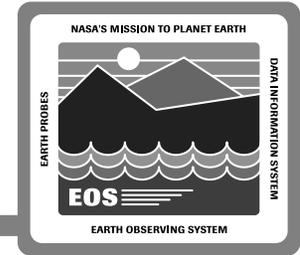
# AHWGP Required Capacity vs. Provided Capacity\*



DAAC	AHWGP Required Capacity				Provided Capacity			
	Release	Peak MFLOPs	Peak I/O Bandwidth	Disk Volume	Platform	Peak MFLOPs	Peak I/O Bandwidth	Disk Volume
LaRC	IR-1	1,100	25 MB/sec	30 GB	SMP (4 CPU)	1,200	320 MB/sec	30 GB
	Rel A	+2,333	No Change	No Change	+SMP (8 CPU)	+2,400	+320 MB/sec	No Change
MSFC	IR-1	10	25 MB/sec	5 GB	Uniprocessor WS	125	100 MB/sec	5 GB
	Rel A	No Change	No Change	+5 GB	+Uniprocessor WS	+125	+100 MB/sec	+5 GB
EDC	IR-1	120	25 MB/sec	35 GB	SMP (2 CPU)	600	320 MB/sec	35 GB
	Rel A	+57	+156 MB/sec	+35 GB	No Change	No Change	No Change	+35 GB
GSFC	IR-1	2,150	+250 MB/sec	75 GB	SMP (8 CPU)	2,400	640 MB/sec	75 GB
	Rel A	+2,040	No Change	No Change	+6 CPU	+1,800	No Change	No Change

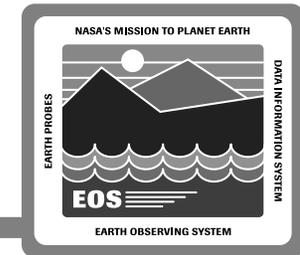
\* Does not include DAO

# Symmetric Multiprocessor (SMP)



- **SMP**
  - is “multiple workstations in a box”
  - all processors are identical
  - provides coherent shared memory
- **Rationale for selection**
  - prototyping studies
  - *environment supports conventional sequential* in addition to symmetric and distributed memory parallel processing
  - best price/performance with more capacity per box for \$\$\$
  - market survey has indicated that future SMP performance is increasing at a faster rate than MPPs (does not preclude MPPs for Release B and beyond)
  - the interconnect between processors has higher bandwidth than many external interconnects

# Preliminary Planning Database Size Estimates



## Assumptions :

- Estimated for CERES production support for one month based on Epoch "c" statistics.
- Estimated Production Requests : average of 40 per month
- Estimated average number of CERES Data Processing Requests generated per day : 200
- Estimates based on number of PGE activations/day for CERES at LaRC (Release A)

## Planning Database Size Estimates:

- Production Requests : 64 Kbytes
- Data Processing Requests : 160 Mbytes
- Plans : 720 Kbytes
- Data Availability Schedule Data : 540 Kbytes
- Static DBMS data : 122 Kbytes

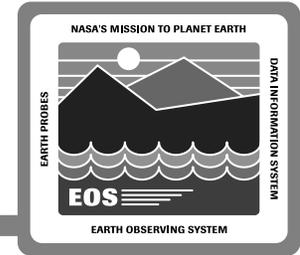
Total estimated Planning (CERES) database size for 1 month : 162 Mbytes\*

\*\* Raw data, no DBMS overhead included

## Planning Database Throughput:

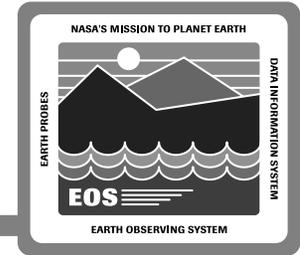
- Background Load
  - Production request activation
  - Plan feedback
  - < 1 access/sec
- Peak Load
  - Plan creation, activation
  - What-if analysis
  - up to 30,000 DB accesses/plan
  - ~100 accesses/sec
  - several minutes for complex plan

# Planning Hardware and Database Sizing



- For IR-1 and Release A (based on preliminary figures of AHWGP data)
  - simple workstation class machine and a commercial DBMS
- Dramatic increase in accesses/sec in Release B calls for
  - high network bandwidth
  - high I/O
- Preliminary estimates for Release B indicate
  - server class machine with robust DBMS may be needed

# Road Map for Planning & Processing Presentation



## Overview

- Concept Drivers, Key Features
- Production Management Flow

## Software Model

## COTS/Prototypes

- Evaluation
- OTS and Software Reuse

## Scenarios

## Cross DAAC Scheduling/Planning

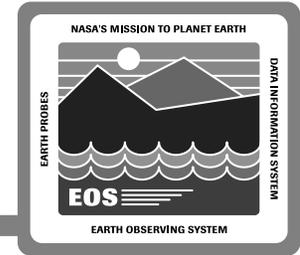
## Other Data Processing CIs

- AI&T Tools
- Science Data Preprocessing

## Hardware

## *Issues*

# Issues



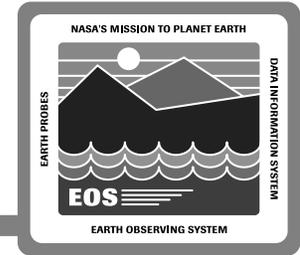
## Data Production Quality Assurance (QA)

- **Statement:** Specifics of DAAC QA yet to be defined by ITs
- **Impact:** Potential for design and/or sizing changes
- **Suggestion:** Work with AHWGP and ESDIS to understand DAAC QA requirements and determine if current 4-level QA infrastructure is adequate

## Dynamic Planning (Replanning)

- **Statement:** Automatic initiation of replanning as a response to changing events which caused some level of deviation from Plan
- **Impact**
  - affects resource optimization
  - influences local DAAC control over data processing
  - affects local DAAC autonomy with ripple effects of replanning
- **Suggestion:** Work with ESDIS and DAACs for a resolution

# Issues (cont.)



## Unpredictable Conditional Activation

- **Statement:** Determine method for supporting the unpredictable portion of PGE activation rule #5, based on metadata values
- **Impact:**
  - resource profile of a PGE is very unpredictable
  - affects predicted resource utilization
- **Suggestion:** Several options under consideration. Working with AHWGP, ESDIS, and DAACs for a resolution and plan it for Release B

## Planning horizon

- **Statement:** Current design supports any planning horizon (e.g. 6 months to a year)
- **Impact:** Increase in DB size, performance requirements, cost
- **Suggestion:** During the CDR phase, we plan to work with ESDIS, DAACs and instrument teams to determine by instrument the planning horizon to be supported