Application of Data Grid Technology to Earth Remote Sensing Science Data Segment Architectures

Samuel D. Gasster, Craig A. Lee
Computer Systems Research Department, M1–102
The Aerospace Corporation
PO Box 92957
Los Angeles, CA 90009–2957
gasster@aero.org
(310) 336-7715

Jeffrey T. Lubelczyk
Science Data Systems Branch
NASA GSFC
Greenbelt, MD 20771

Robert Harberts
Global Science & Technology, Inc.
Greenbelt, MD 20771
This work is supported by NASA/GSFC.

Outline

<u>Topics</u>

- Background and Motivation
- Requirements and Operations Concept
- Why Use Grid Technology?
- The Advanced Data Grid (ADG) Prototype
- ADG Development Plans
- ADG Architecture
- ADG Implementation
- ADG Test Plans
- ADG Schedule
- Summary and Status

Background and Motivation

The National Polar-Orbiting Operational Environmental Satellite System

- The NPOESS Preparatory Project (NPP) will launch a satellite in late 2006
- Risk Reduction for NPOESS: sensors, algorithms, ground data processing
- Requirements for a ground system to support climate scientists

The Science Data Segment (SDS) for NPP

- SDS must capture lifetime Science Mission Data from NPP
- SDS must support science teams distributed across North America

Science Mission Data Rates and Volumes

- Increased Data Rates (>10Mbps) from Remote Sensing and Earth Science missions
- Overall data volume for a typical 5 year mission duration > 1 peta-Byte (10^{15} Bytes)
- Data Reprocessing Requirements can result in throughput as high as 20X real-time

Background and Motivation

Science Data Workflow Management

- **Management and optimization of multiple data processing streams**
- **Increased need to provide data discovery and access to large, geographically distributed user community**
  - **Data discovery becoming more complex; need a solution that is scalable and will interoperate in a highly heterogeneous resource environment**
  - **Value added products and models are becoming more computationally intensive and data assimilation is becoming mainstream; e.g., climate models that use direct assimilation of satellite data**
- **Provide long term archive and retrieval of critical measurement data and make this data available for reprocessing**

Identified Science User Needs

- **Transparent, Secure and Timely Access to Data**
- **Sharing of Resources: Data, CPUs, Software: Tools, Algorithms, Models**

Requirements and OpsCon

Requirements

- Project Goals and Requirements were drawn from several sources
  - NASA NPP Requirements provided examples of a near term system that will generate peta-bytes of data over 5 year mission life
  - Other studies by NASA and NRC for Earth Science Data Systems (see references)

Operations Concept (OPSCON)

- To further specify the prototype we developed several operations concept “scenarios”
  - Scenarios describe how the system will be used; Science Data Segment workflow and data life cycles
  - Based on NPP OpsCon and current NASA Data Archives
  - *Mimic* typical scenarios for future remote sensing and Earth science ground systems
- Used to define the prototype architecture
  - Map these requirements into a Grid Architecture based on defined services and derived resource requirements
  - Architecture should also support testing

Requirements

Example Requirements

- Data Ingest and Processing (raw and derived products)
- Provide access to derived data products from on-line mission storage
- Provide access to derived data products from long term archival storage
- Provide search/retrieval access to mission storage for ~20 science team members
- Provide Near-line Storage for ~2.6 pB (5 year mission life)

Data Processing Analysis

REAL-TIME PROCESSING				
Data Type	GB/Day	Data Rate (MB/sec)	Data Rate + 25% Margin (MB/sec)	2 Days of Data (GB)
L0	300	3.47	4.34	750
L1A	259.2	3.00	3.75	648
L1B	451.2	5.22	6.53	1128
L2	23.3	0.27	0.34	58.25
L3	98	1.13	1.42	245
TEU	15	0.17	0.22	37.5
Ancillary Data	5	0.06	0.07	12.5
	1151.7		16.66	2879.25
			5 Year Mission Life:	2.629 pB
REPROCESSING				
Reprocessing Ratio =	20			
Reprocessing Data Rate (MB/sec)	333.25			
Data Volume (2 Days) (TB)	57.585			

Operations Concept

Studied Several Operations Concepts for an SDS

- Provide Basis for defining test scenarios
- Relate Directly to Actual SDS Workflow
- Use Limited Total Data Volume

Example OpsCon Scenarios

- Level 1 Data Ingest (ingest of calibrated satellite data)
- Cal/Val Data Processing
- Climate Data Processing
- General Science User Data Discovery and Manipulation
- Data Reprocessing
- Details provided later in briefing

Data Life Cycle Simulations

- Test cases defined to simulate typical operations
- Cover Limited Time Spans: 1, 5 and 10 Days
- Match Data Rates and Data volumes (with replay)

Why Use Grid Technology?

Grid Technology Context

- Grid Technology provides a new approach to the development of distributed data and compute intensive ground system architectures
- NASA and other large scientific organizations have started development of various types of distributed data processing and storage/retrieval systems using Grid technology
  - NASA/Ames Research Center/Information Power Grid: Aerospace Applications
  - Particle Physics Projects, Earth System Modeling, Digital Sky Survey

Grids Provide High-level Design Tools For Distributed Systems

- Defines coherent Grid Architecture, Protocols, and APIs
- Integrates Information Discovery, Metadata Schemas and Ontologies
- Integrates Resource Discovery and Management
- Integrates Security Mechanisms
- Supports Dynamic Virtual Organizations
- Headed for standardization in GGF, www.ggf.org

Why Use Grid Technology?

Benefits Of Grid Technology For Science Data Systems

- Open, Extensible Systems
  - Supports distributed teams and resources
  - Utilizes common algorithms or standard software tools
  - Avoids stove-piped, closed systems
- Manage Cost and Growth
  - Don't need to buy everything up-front; grow as you go
  - Leverages available resources
  - Utilizes high performance commodity components; redundancy is relatively cheap

Benefits Of Grid Technology For Scientists

- Enables a Core Set of Application-Specific Services to Support:
  - Data discovery: Easily find what you are looking for
  - Data retrieval: Move the data to where you need it
  - Data processing: Standard tools, common algorithms, shared processing power

The Advanced Data Grid (ADG) Prototype

Project Goals

- **Develop a prototype Earth Science Data Segment using Grid Technology**
  - **Compatibility with Existing Metadata Standards and Data Sets**
  - **Suitability for Processing Scenarios, e.g., Filtering, Workflows, etc.**
  - **Suitability for Scientist Collaborations, e.g. Virtual Organizations**
- **Assess scalability and sizing of Grid Technology for Earth Science Data Segment**
  - **Assess performance of specific implementation and identify limitations in various layers of the Grid architecture**
  - **Experiment with different approaches to overcome these limitations**

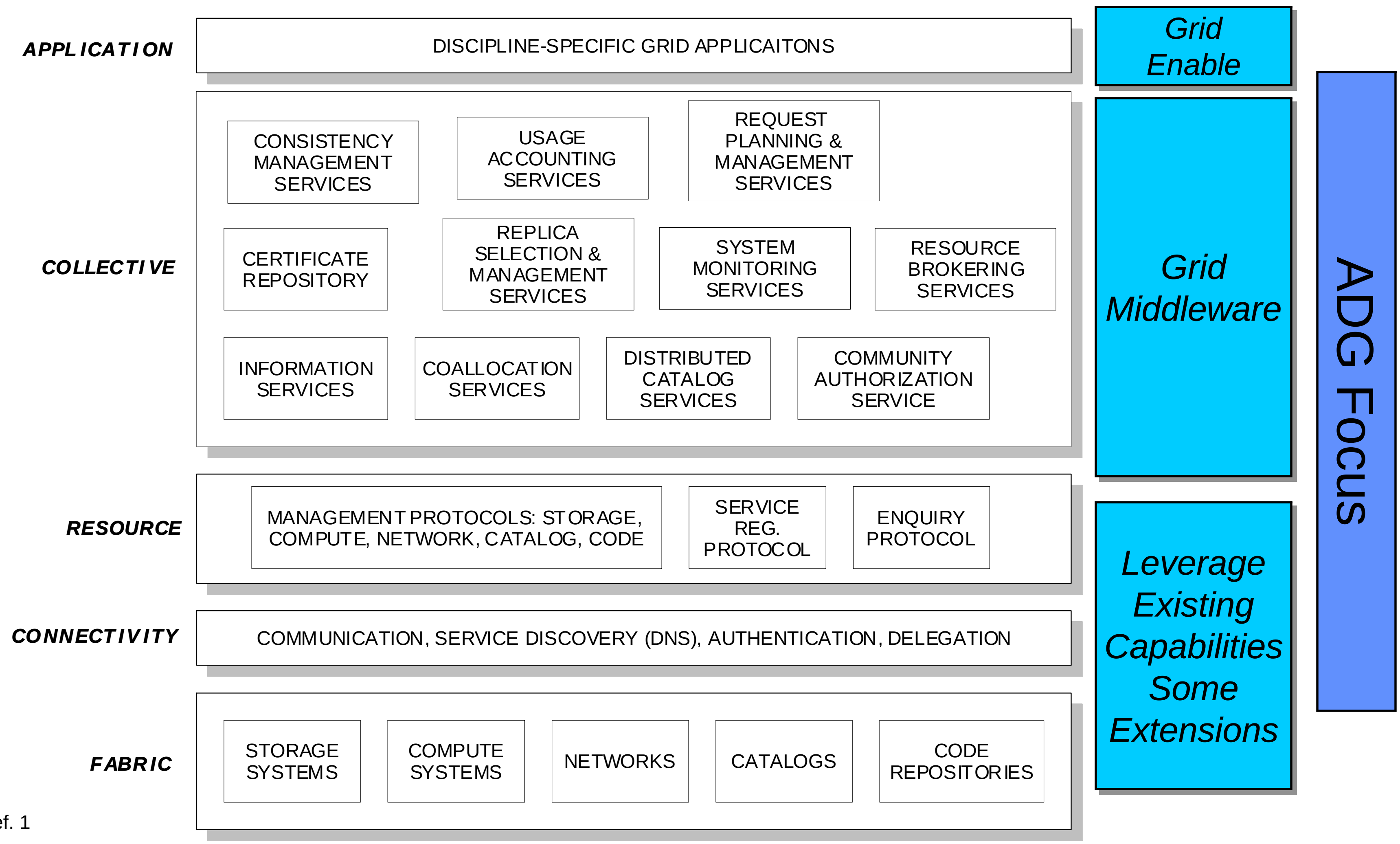
ADG Prototype Development Plans

- **Develop top-level prototype Requirements and Operational Concept**
- **Define ADG Architecture based on *ADG Grid Service Model***
- **Define specific prototype design based on this architecture and constrained by:**
  - **Resource availability**
  - **Project Duration**
  - **Cost**
- ***Adopt NASA Information Power Grid and Globus Toolkit 2.4 as implementation vehicle***
- **Perform prototype testing and performance evaluation**
- **Report Results**

ADG Prototype Project goal is to address sizing, performance and scalability of grid technology for a peta-byte class Earth Science ground system.

Grid Architecture Layer Diagram

–"…Major Data Grid Reference Architecture elements, showing how they relate to other Grid services. Shading indicates some of the elements that must be developed specifically to support Data Grids." – From Ref. 1



After Ref. 1

ADG Service Model

<u>ADG defines services for:</u>

- **Process and Resource Management/Monitoring**
- **Ingest: any data type including Satellite and Cal/Val related data**
- **Export: to external archives**
- **Mission storage and ancillary data management: Metadata Catalog Services, Replica Location and Storage Resource Management (retrieval)**
- **Data Filtering Services (sub-setting, aggregation, gridding)**
- **Data Processing and Reprocessing Services**
- **Data Subscription and Notification**
- **Workflow for each area**
- **Science Model Implement Services (e.g., Numerical Weather Prediction, Radiative Transfer)**

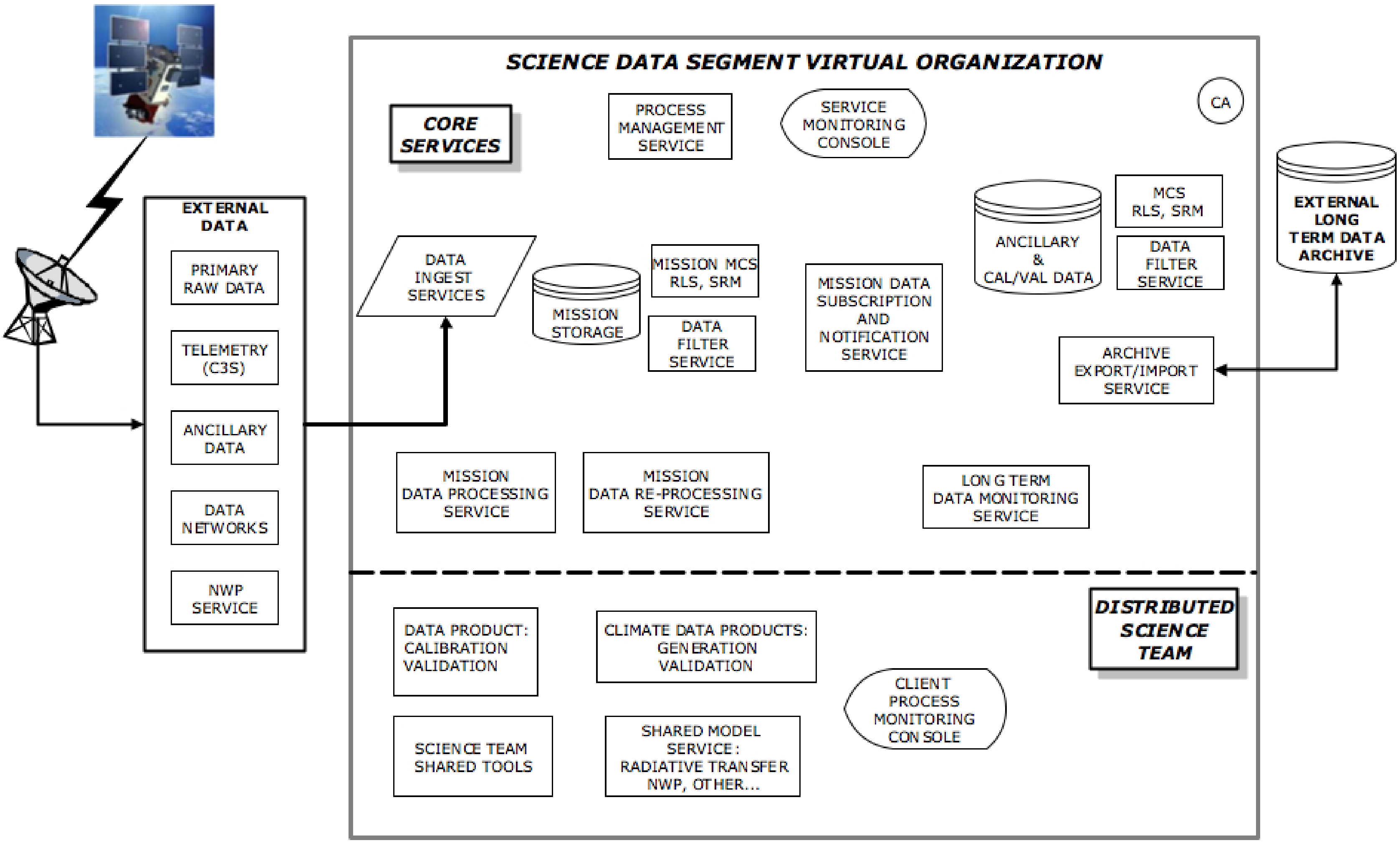
<u>Use Basic Building Blocks</u>

- **See Next Chart**

<u>Services many be instantiated by ADG team or by Science Team member(s)</u>

- **Grid services provide by Core as well as the Science Team members**
- **Example:**
  - **Sensor calibration team may define a set of services to support their sensor calibration activities**
- **A Climate Data Product (CDP) team may define a service for their specific CDP**

ADG: Grid Services View



Notional grid services architecture that support SDS-like requirements

NPP SDS Functions and Grid Services

Develop An Implementation of the NPP SDS Functions

- Climate Data Calibration Function
- Distributed Climate Data Product Algorithm Validation Function
- Climate Analysis and Research Function
- Instrument and Algorithm Calibration and Validation Function

Common Elements Mapped to Grid Services

- Science Team members provide
  - Basic generation of the required data products
  - Product calibration/validation
  - Algorithms, models and associated software
  - Local versions of: MCS, RLS, SRM, Data Filtering (to manage own and team shared data)
  - Provide domain specific tools that they desire to share

- Core Grid Services
  - Data Management: Ingest, MCS/RLS/SRM, Data Filtering
  - Data Processing: Mission data processing and re-processing (key driver!)
  - Subscription/Notification
  - Numerical Weather Prediction Model Service
  - Export to long term data archive

NPP SDS Functions and Grid Services

Allow users access to the resources they need to perform their tasks

- **Don't duplicate the same capabilities at multiple sites**
  - **Share them with other grid members**
- **Allow users access to common resources**
  - **NWP and RT models, Data filtering (sub-setting)**
  - **All members using the current versions**

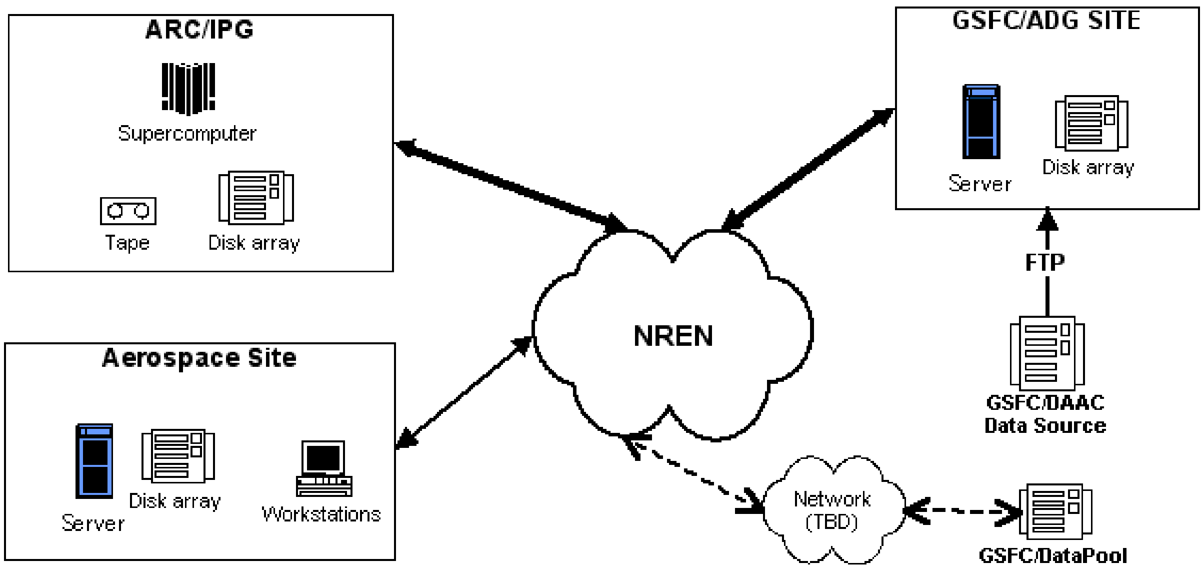
Grid architecture naturally supports distributed resources and services

- **Supports a distributed Science Team with resource sharing**
- **Some capabilities, e.g., primary mission storage, provided by a more centralized facility to support operational processing/re-processing requirements**
- **Distributed team members easily accommodated with grid service model**

ADG PHYSICAL VIEW

ADG Certificate Authority
Computing Resources
Tape Archive

Primary On-line Data Store
And MCS Site



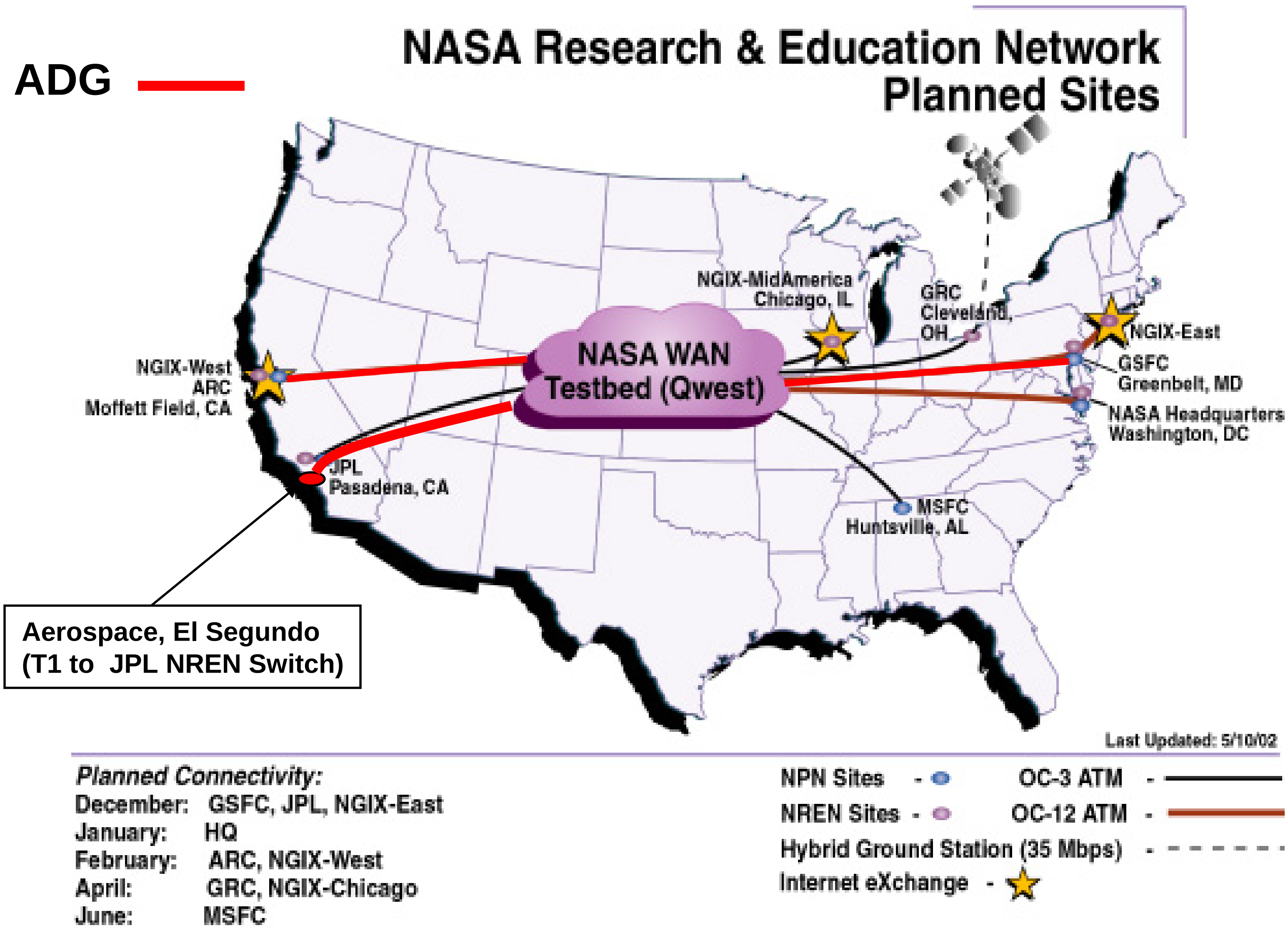
Science User Simulation
Application Development and Testing
Federated MCS

Level 1 data processing site
Grid-enabled access via Data Pool prototype

<u>All sites include</u>

– **Common grid S/W: Globus 2.4, SRM Clients, etc.**
– **Performance monitoring tools**

Hosting the ADG on NREN



Prototype Implementation

Implementation

- Focus on Science User perspective
- Leverage Existing Tools Whenever Possible (e.g., GTK)
- Demonstrate core data management and manipulation functions
  - MCS (development), RLS (GTK), File Transfer (GridFTP, RFT, HTTP)
  - Metadata Schema
- Resource Sharing: Algorithms and Data Processing
  - Data Filtering: DataCutter
  - Shared Algorithm or Tools
    - Radiative Transfer Models
    - Numerical Weather Prediction Models
    - Calibration Algorithms
    - Climate Data Products
- Metadata Catalogs and Storage Resource Management
  - Investigated Several Approaches
  - SDSC SRB/MCAT
  - Globus MCS, RLS
  - Evaluating New Approach to MCS that allows federation of catalogs

ADG Test Plans

Developmental Testing

- Continuous approach to testing; test as you go
- Test within a “layer” (infrastructure testing and benchmarks)
- Test across a “layer” (interoperability, performance, scalability)

OpsCon Testing

- Testing Driven by requirements and scenarios (e.g., data life cycles)
- Level 1 Data Ingest
  - Data Pool site is treated as the ADG Level 1 processing site with only short term data storage
  - Ingest L1 data from Data Pool short term storage to ADG On-line mission storage [GSFC/ADG]
  - Transfer L1 data from on-line mission storage to archive storage [ARC/IPG]

ADG Test Plans (cont.)

Testing Driven by Requirements and OpsCon Scenarios (cont)

- **Cal/Val Data Processing**
  - **Cal/Val Process collects satellite data and collocated in-situ measurement for the Calibration and Validation of various data products**
  - **Ingest Calibration Level 1 data from Data Pool; Simulate Field Campaign in-situ data [GSFC/ADG]**
  - **Aerospace Site will act as Science User performing Cal/Val functions**
    - Request sub-set of Calibration Level 1 data that matches up (spatial/temporal) with field campaign in-situ data
    - Transfer sub-set L1 data and field data from mission storage [GSFC/ADG] to Aerospace working data cache
- **Climate Application Processing**
  - **Demonstrate an actual application using MODIS data to generate a climate data product (CDP)**
  - **Aerospace Site will act as Science user performing CDP generation functions using ADG resource**
    - Develop CDP application software that can be run on available ADG compute resources
    - CDP Application will request transfer of required data to data cache collocated with compute resources
    - CDP Application results will be transferred back to mission storage [GSFC/ADG] and to archive storage [ARC/IPG]

ADG Implementation Schedule

Project Implementation Schedule

- **Four Phases over 3 years**
- **Phase I: Initialization**
  - Install and test H/W and S/W at each site, as required
  - Local and some interface testing
  - Security implementation; Certificates issued
- **Phase II: Baseline**
  - Testing of basic Grid interoperability
  - Grid Security Testing
  - ADG Prototype Benchmark testing
  - Data acquisition
- **Phase III: Grid Testing**
  - Full-up Test and Measurement of the ADG
  - Exercise Science Data Segment Workflow
- **Phase IV: Application Demonstration**
  - Continued grid testing
  - Demonstration of a grid enabled science application

Status and Future Potential

STATUS

- ADG is currently in Phase I: Initialization
- ADG Working Group formed from member organizations
  - GSFC, ARC/IPG, NREN, Aerospace
- Hardware installed at GSFC/ADG Site; working on configuration and baseline testing
- Working towards baseline MCS installation/configuration/testing
- Data Pool capability expected sometime in 2004
- Aerospace working towards NREN connectivity
- Funding Terminated in 4th Qtr 2003, working to get revised funding

LESSONS LEARNED TO-DATE

- Cross Organization Resource Sharing
- Networking: IP address space, firewalls, routing
- Metadata Catalog Issues
  - Difficulty implementing large complex earth remote sensing schemas for metadata in existing systems
  - Lack of federation capability; MCS becomes a single point failure for an operational system

References

1. *A Data Grid Reference Architecture* (Draft of February 1, 2001), Ian Foster, Carl Kesselman (see http://www.ppdg.net/archives/ppdg/2001/doc00013.doc)
2. Global Grid Forum: http://www.gridforum.org/
3. NPP Web Site: http://jointmission.gsfc.nasa.gov/
4. Globus Web Site: http://www.globus.org/
5. Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites, Committee on Earth Studies Space Studies Board Division on Engineering and Physical Sciences, National Research Council (2000). See http://www.nas.edu/ssb/cdmenu.htm

Acronyms

ADG	Advanced Data Grid
ARC	Ames Research Center
CDP	Climate Data Product
CDR	Climate Data Record
CEOS	Committee on Earth Observation Satellites
EDR	Environmental Data Record
GGF	Global Grid Forum
GSFC	Goddard Space Flight Center
GT	Grid Technology
GTK	Globus Toolkit
H/W	Hardware
IPG	Information Power Grid
MCAT	Metadata Catalog
MCS	Metadata Catalog Service
NASA	National Aeronautics and Space Administration
NPOESS	National Polar-orbiting Operational Environmental Satellite System
NPP	NPOESS Preparatory Project
NRC	National research Council
NREN	NASA Research and Education Network
NWP	Numerical Weather Prediction
OGSA	Open Grid Services Architecture
QoS	Quality of Service
RLS	Replica Location Service
RTM	Radiative Transfer Model
S/W	Software
SDS	Science Data Segment
SRB	Storage Resource Broker
SRM	Storage Resource Manager