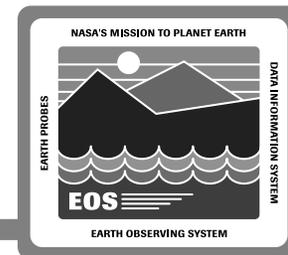


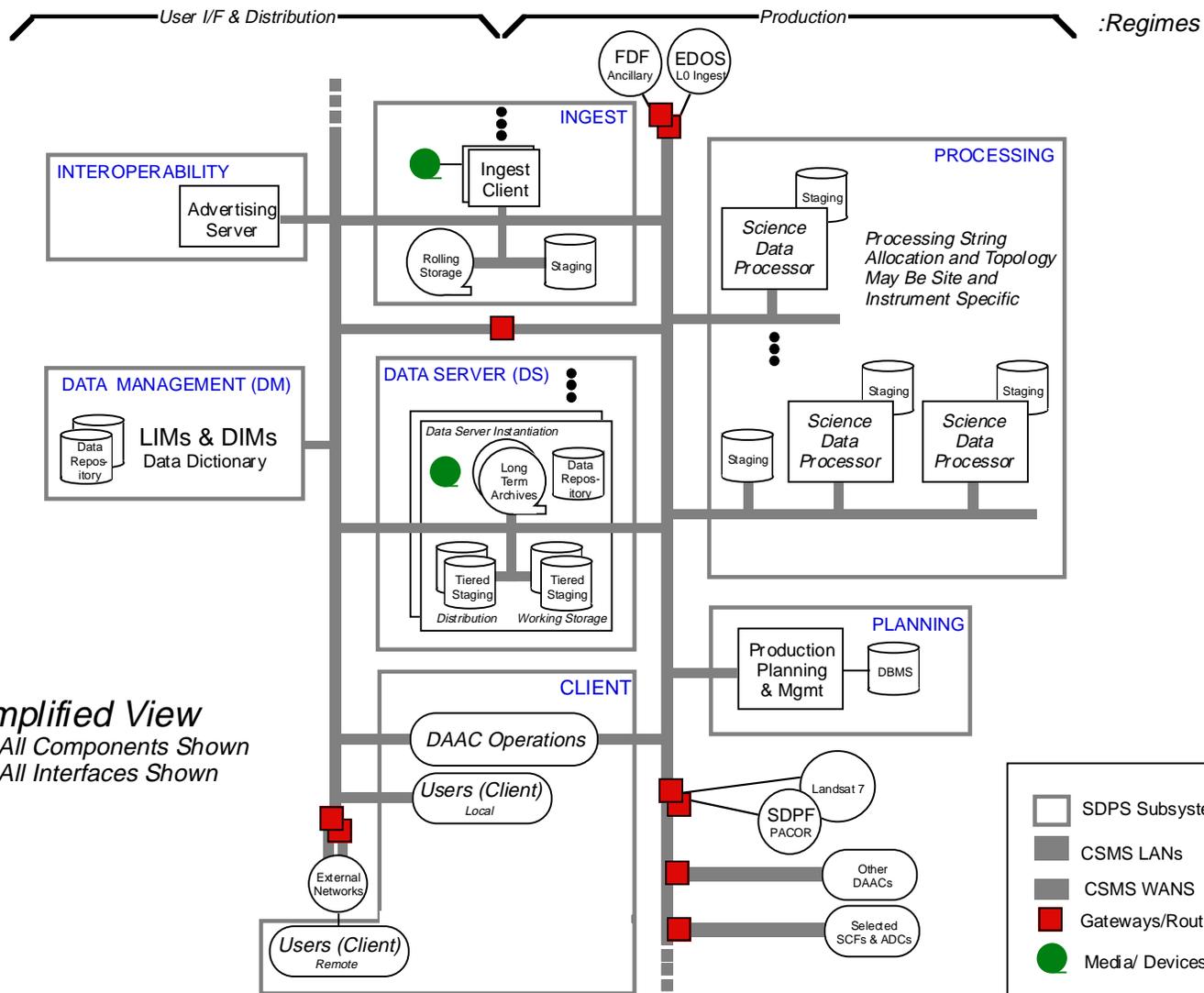
SDPS Hardware Implementation

Eric Dodge/Mark Huber

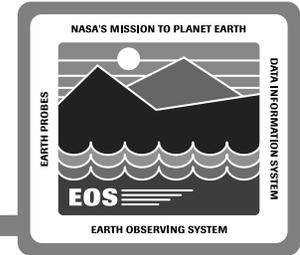
System Design Review - 28 June 1994



Hardware Design Overview



SDPS Implementation Architecture Overview

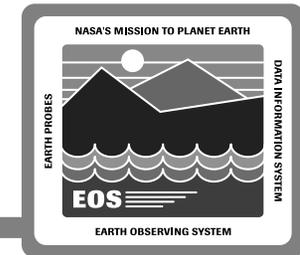


See Next Page



ADDITIONAL MATERIAL

System Issues / Strategies & Features



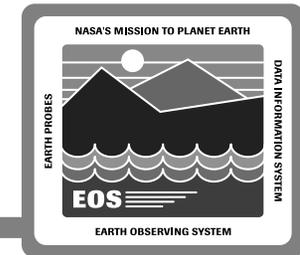
Issue: Large user access concurrent with production demands

- **Strategy: Selective partitioning of system resources**
 - **Division of framework into User/Distribution & Production regimes**
 - **Minimize contention: Inter-Disciplinary Science (IDS) flows, subscriptions from product dependencies and production**
 - **Partitioned “pools” of staging and other hardware resources**

Issue: Evolvability

- **Strategy: Use commodity COTS utilizing open/recognized standards & partition system to allow for technology insertion**
 - **No custom hardware**
 - **Use of open or strategic recognized / emerging standards**
 - **Strategic partitioning for technology insertion minimizing global impacts**

System Issues / Strategies & Features (continued)



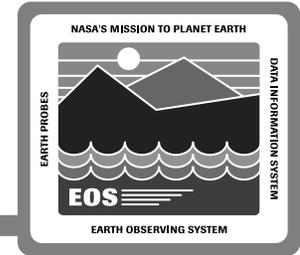
Issue: Large I/O streams

- **Strategy: Numerous -- High speed technologies, spread of I/O across component instantiations, subpools of staging.**
 - Multi-tiered LANs/WANs
 - Dedicated subnetworks applied where and as needed (localization)
 - Shared subpools of RAID staging
 - Network attached peripheral solutions to minimize data “hops”
 - Separation of control from prime data flow (large DAACs)

Issue: Large staging demands (e.g. product dependencies, dist., etc.)

- **Strategy: Numerous -- multi-levels of technologies, pools of staging**
 - Multi-tiered staging / virtualization (minimize RAID)
 - Intelligent / pre-staging of dependencies

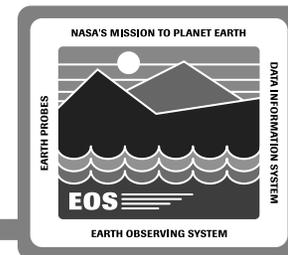
System Issues / Strategies & Features (continued)



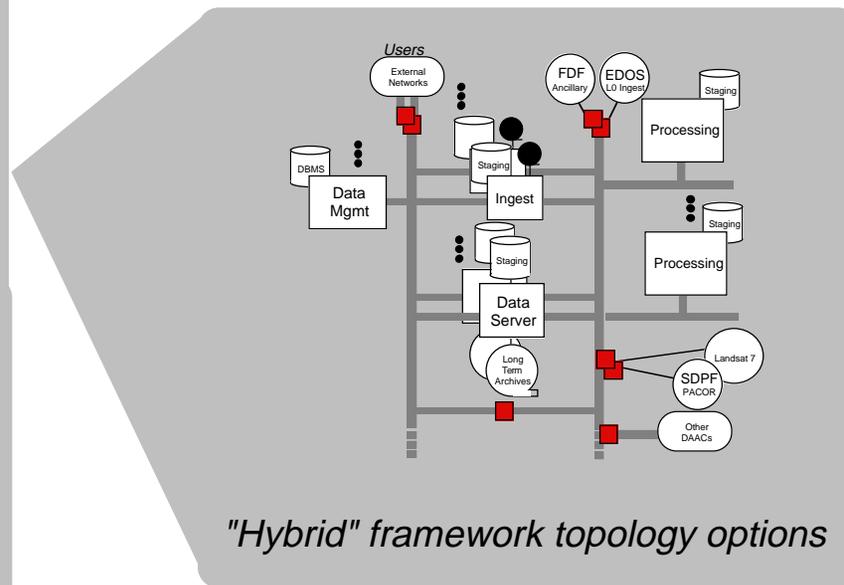
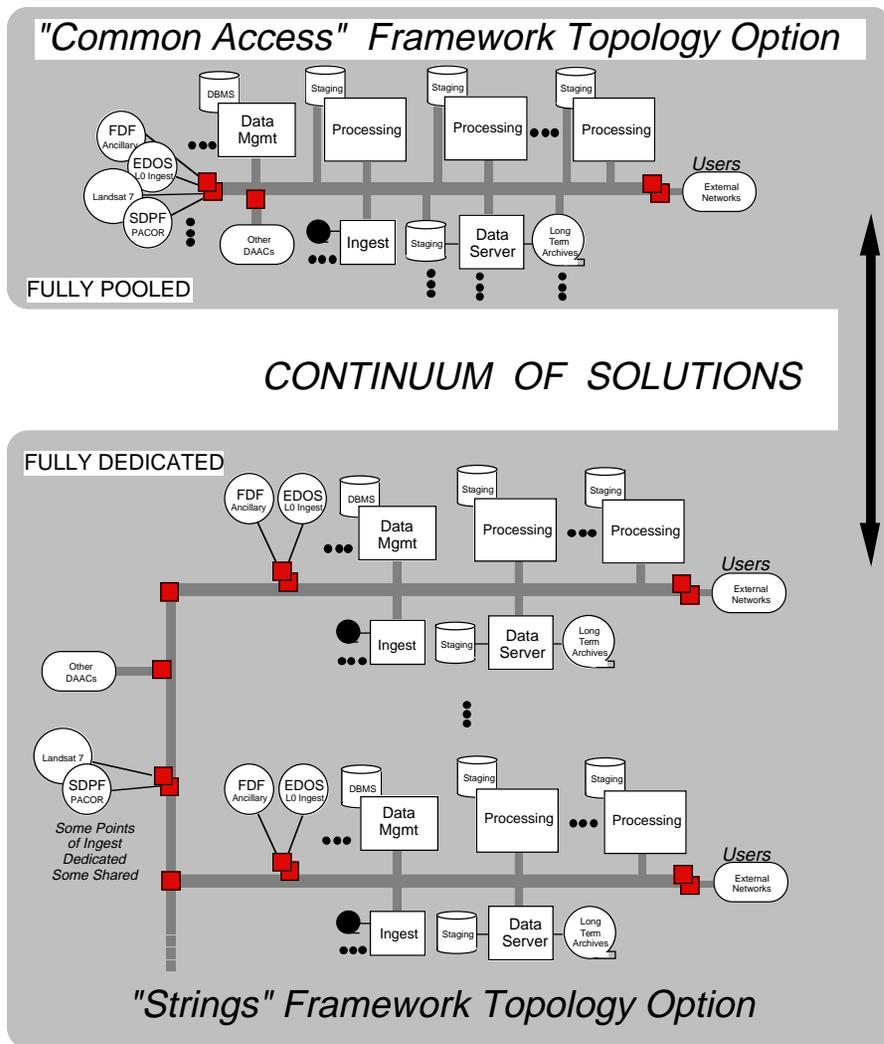
Issue: Need for reliable operations

- **Strategy: Numerous -- redundancy, technology selection, resource pooling and partitioning**
 - **Redundancy: processing (test & backup), planning**
 - **Hot standby: ingest clients (e.g. L0)**
 - **Fault resistant/tolerant technologies: RAID staging, archive, networks**
 - **No single network point of failure (for I/O paths with high RMA)**
 - **Pools of resources for graceful degradation: Data Server, Processing**

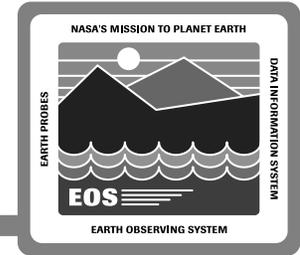
Issue: Scalability (addressed throughout)



Topology - Hybrid Solution(s)



Topology Selection



“Selection” based on Modeling, Cost and Design Analysis

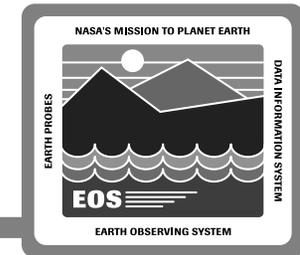
Key findings

- **Topology effects are relatively cost insensitive**
 - Hybrid “wins” (slim margin under Strings, 7% under Common)
- **Selection made based on evolvability and scalability issues**
 - Site uniqueness need
- **Hybrid is a family of solutions between extremes**

Major contributing disadvantages

- **Common Access: large central staging pools magnify contention, growth & insertion impacts global resources throughout, no tangible distribution and production isolation**
- **Strings: Data Server dedication, no tangible distribution and production isolation**

Hybrid Topology Implications On Architecture



Not a single topology but a family of permutations based on site needs

Resources are shared or dedicated to tasks as needed

Inherent separation of user/distribution from routine production activity

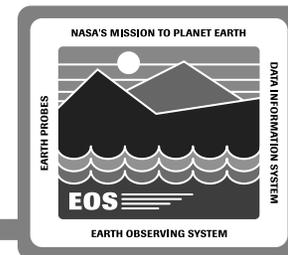
- **Production: product dependency flows, L0 ingest, QA, etc.**
- **User/Dist: subscriptions, media distribution, electronic access**
- **Service exceptions: adhoc processing, subsetting, browse, RST**

Pools of processing resources (chains/strings)

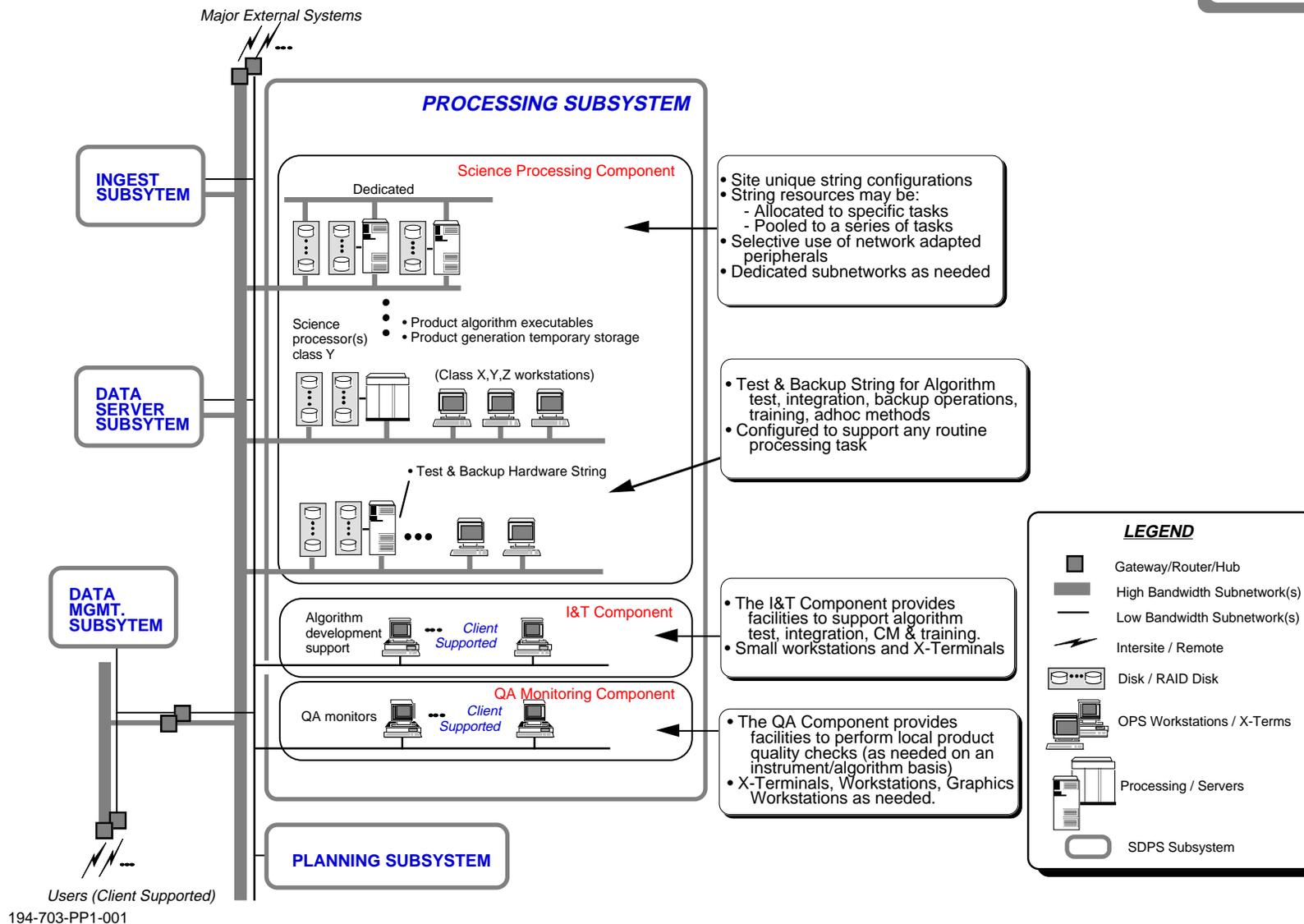
- **On demand (production or I&T strings)**
- **Routine (production strings)**
- **Adhoc (I&T strings)**

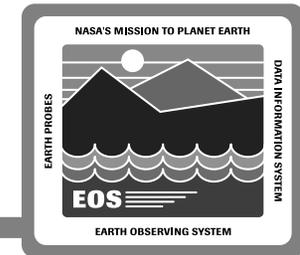
Sharing of other resources as required at the DAAC site

- **DM, DS, Ingest Clients**
- **Inter-center communications, major external interfaces (e.g. L0)**



Processing Subsystem





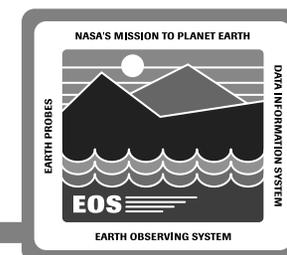
Processing Issues / Features

| Major Subsystem Issue Description | Primary H/W components | Major Alternatives | Supporting Design Features |
|---|---------------------------------|--|--|
| Diverse and changing science algorithm processing requirements | Science Processing | Pooled or dedicated hardware approaches | <ul style="list-style-type: none"> • Processing "strings" (chains) a mix of dedication and pooling • Routine production: tuned strings • Adhoc: planned & queued. |
| High bandwidth I/O associated with large numbers of product dependencies | Science Processing | Host or network attached staging with high bandwidth LAN / WAN communications | <ul style="list-style-type: none"> • Large strings: Network attached RAID (avoid excessive I/O thru hosts) • Separate control and data paths • Dedicated & internal subnetworks • Semi-dedicated pools of staging. • Mix of host and network staging. |
| Algorithm integration and test activity is required in parallel to ongoing routine production | Science Processing I&T | Pooled or dedicated hardware approaches | <ul style="list-style-type: none"> • One or more test & backup strings • Site dependent configuration • I&T, training, adhoc/on-demand, subsetting/browse support |
| Future requirements changes or growth in processing requirements that dictate the need for distributed processing | Science Processing | MPPs, parallel concurrent processing, SMP, etc. * | <ul style="list-style-type: none"> • Strings allow for augmentation and / or insertion of new technologies without global impacts. • Algorithm dependent not architecture dependent. |
| Large I/O and processing requirements require constant operational load leveling, coordination and planning | n/a (See Planning Subsystem) | Distributed or centralized production load "scheduling" (referred to as queuing) | Production load leveling, stusing, and queuing is provided in a shared capacity with the Planning Subsystem. |

* Use of MPPs, SMPs, is algorithm dependent (efficiency issue)

Release B Processing Sizing

(SDR Baseline *1 Volume & *1 Processing)



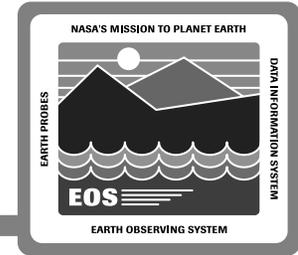
| Site | String | Processing Class | Total # of CPUs | Capacity (Proc.) (MFLOPS) | Staging Type | Capacity (Staging) (GB) | Capacity (I/O) (MBps) | Comm. Classes |
|-------|--------|--|-----------------|---------------------------|-----------------|-------------------------|-----------------------|-----------------------------|
| ASF | none | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| EDC | ASTER | 1 PP (14) | 14 | 4200 | RAID net. adapt | 566.6 | 640 | FC / HiPPI FDDI 802.3 |
| | MODIS | 1 UWS (1) | 1 | 300 | Host adapt | 14.7 | 132 | FDDI 802.3 |
| GSFC | MODIS | •6 VS (16) •1 VS (8) | 104 | 20800 | RAID net. adapt | 3712.0 | 8320 | FC / HiPPI FDDI 802.3 |
| JPL | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| LaRC | CERES | •3 PP (4) •1 PP (3) (or SMP W/S) | 15 | 4500 | RAID net. adapt | 472.1 | 528 | FC / HiPPI FDDI 802.3 |
| | MISR | •3 VS (16) •1 VS (12) | 60 | 12000 | RAID net. adapt | 638.1 | 4800 | FC / HiPPI FDDI 802.3 |
| | MOPITT | 1 PP (3) | 3 | 900 | Host adapt | 9.1 | 132 | FDDI 802.3 |
| MSFC | LIS | 1 UWS (1) | 1 | 300 | Host adapt | 7.0 | 132 | FDDI 802.3 |
| NSIDC | MODIS | 1 UWS (1) | 1 | 300 | Host adapt | 1.8 | 132 | FDDI 802.3 |
| ORNL | none | n/a | n/a | n/a | n/a | n/a | n/a | n/a |

Processor Classes

“VS” = Vector Super-Computer

“UWS” = Uni-processor W/S

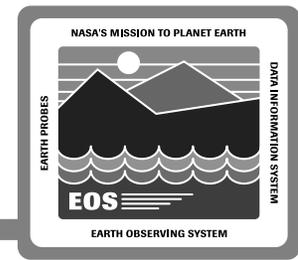
“PP” = Parallel Processor
Includes SMP hosts &
all forms of Multi-
processors



Processing Scalability Strategies

Scalability strategies for the Processing Subsystem

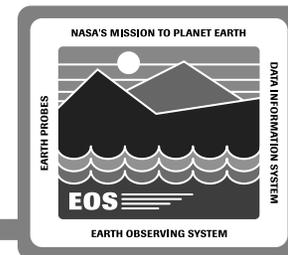
- Addition of new processing strings/chains
- Addition of CPUs to existing host computers (tech. dependent)
- Addition of hosts to existing strings/chains (subnetwork loading)
- Addition of dedicated/new subnetworks (inter/intra string I/O)
- Augmentation of existing hosts with increased I/O subsystem bandwidth / communications (tech. dependent)
- Technology refresh (augmentation and/or replacement) or “graduation” to next technology plateau
- Addition/augmentation of staging resources
- Combination approaches



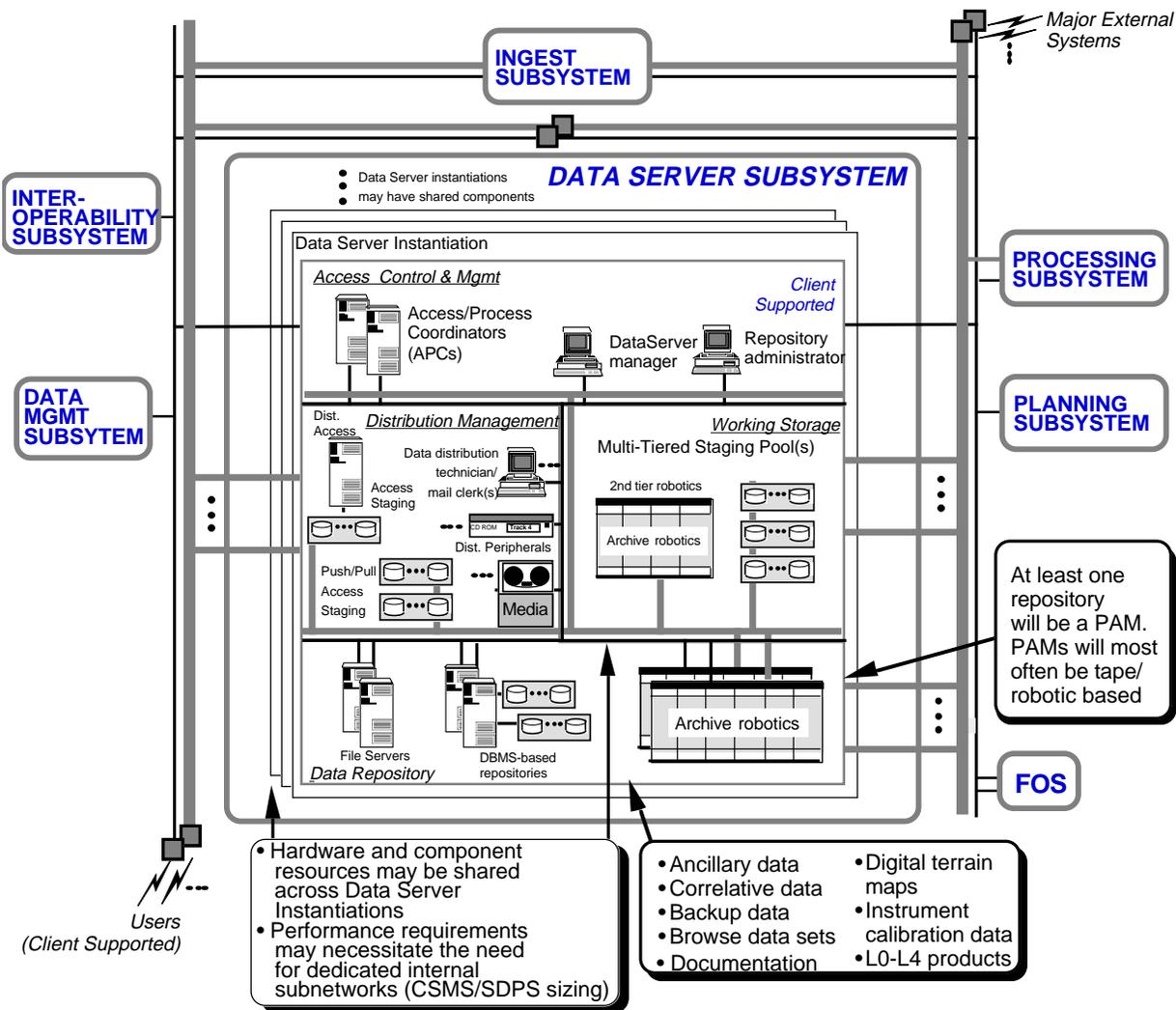
Processing Scalability

(For LaRC/NSIDC - *2 Volume & *8 Processing)

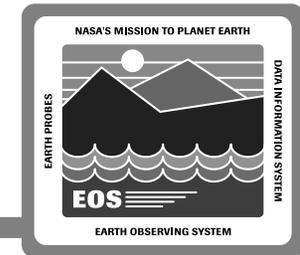
| Site | "String" | Proc. Percent Increase (% avgs) | I/O Percent Increase (% avgs) | Staging Percent Increase (%avgs) | Processing Scalability Strategies That Would be Required For 2 * Storage Size & 8 * Processing (Release B) |
|-------|----------|---------------------------------|-------------------------------|----------------------------------|--|
| LaRC | CERES | 778% | 200% | 1467% | <ul style="list-style-type: none"> • Possible introduction of additional strings. • Migration to higher capacity processing class (from multi-CPU PP to large VS with multi-CPU) • Additional pooled RAID storage required. • Addition of HiPPI/FC paths to storage (dedicated subnetwork support). • No significant changes to intra-string I/O. |
| | MISR | 779% | 200% | 1409% | <ul style="list-style-type: none"> • Possible introduction of additional strings with air-cooled VS class processing. • Possible option to migrate to liquid-cooled VS on same or additional string. • Additional pooled RAID storage required. • Addition of HiPPI/FC paths to storage (dedicated subnetwork support). • No significant changes to intra-string I/O. • Definite target for technology insertion and refresh for processing. |
| | MOPITT | 764% | 200% | 1515% | <ul style="list-style-type: none"> • Additional host of same class needed. • Additional host attached staging needed. |
| NSIDC | MODIS | 688% | n/a < | n/a < | <ul style="list-style-type: none"> • No change required in I/O or processing. • Minor disk growth. |



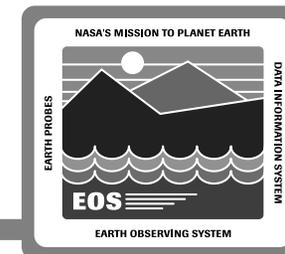
Data Server Subsystem



Data Server Issues / Features



| Issue Description | Design Feature |
|--|--|
| High Internal I/O Rate Required | Use of Network Attached Storage and multiple concurrent I/O channels lead to a high performance, scalable architecture |
| High Performance Data Retrievals on an extremely large Archive data volume | Use of heterogeneous data storage technologies allow for storing data logically and physically so as to optimize performance |
| Data-is-Data | Data, regardless of its storage method, is accessed using the same methods and interfaces. |
| Minimize Magnetic Disk Requirements (Cost) | Use of Network Attached Storage and sharing of disk resources between hardware components lowers the overall magnetic disk requirements for the system |



Data Server Sizing

| Site | WS 1st Tier Staging (GB) | Staging Disk Channels | WS 2nd Tier Drives (MO) | WS 2nd Tier Robotics | PAM Drives | PAM Robotics | Dist Staging (GB) | Access Staging (GB) | File Servers |
|--------|--------------------------|-----------------------|-------------------------|----------------------|------------|--------------|-------------------|---------------------|--------------|
| ASF * | TBD | TBD | TBD | TBD | TBD | TBD | TBD | TBD | TBD |
| EDC | 676 | 8 | 5 | 2 | 13 | 3 (Large) | 761 | 69 | 4 |
| GSFC | 1,388 | 11 | 10 | 3 | 25 | 6 (Large) | 872 | 79 | 5 |
| JPL * | TBD | TBD | TBD | TBD | TBD | TBD | TBD | TBD | TBD |
| LaRC | 581 | 4 | 1 | 1 | 3 | 1 (Large) | 597 | 54 | 1 |
| MSFC | 33 | 1 | 1 | 1 | 1 | 1 (Small) | 45 | 4 | 1 |
| NSIDC | 56 | 1 | 1 | 1 | 4 | 1 (Small) | 77 | 7 | 1 |
| ORNL * | TBD | TBD | TBD | TBD | TBD | TBD | TBD | TBD | TBD |

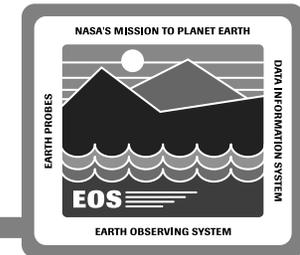
ASF, JPL & ORNL

Release B product set and data support requirements still being defined. Assumed to be small in comparison

Permanent Archive Management (PAM)

PAM Robotics LARGE = 150 TB Capacity per
 PAM Robotics SMALL = 25 TB Capacity per

Data Server Scalability Strategies



Increasing Disk Storage in Distribution and Working Storage

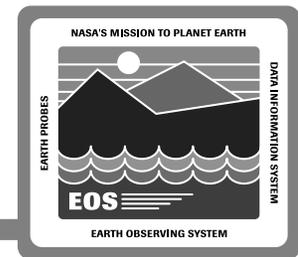
- Network Attached Disk facilitates easy additions to pools
- Network segmentation used when network limited

Segmentation of Repositories

- Mixing of storage solutions allows for technology insertion
- Segmentation allows for expansion (number & type)
- Additional file servers and/or physical repositories

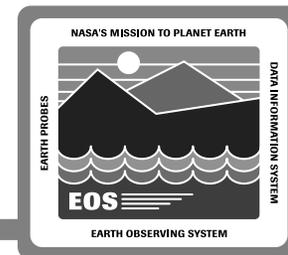
Augment Tiers in Working Storage and PAM

- Larger Working Storage increases/maintain "hit" rate
- Tailor storage repositories to data types

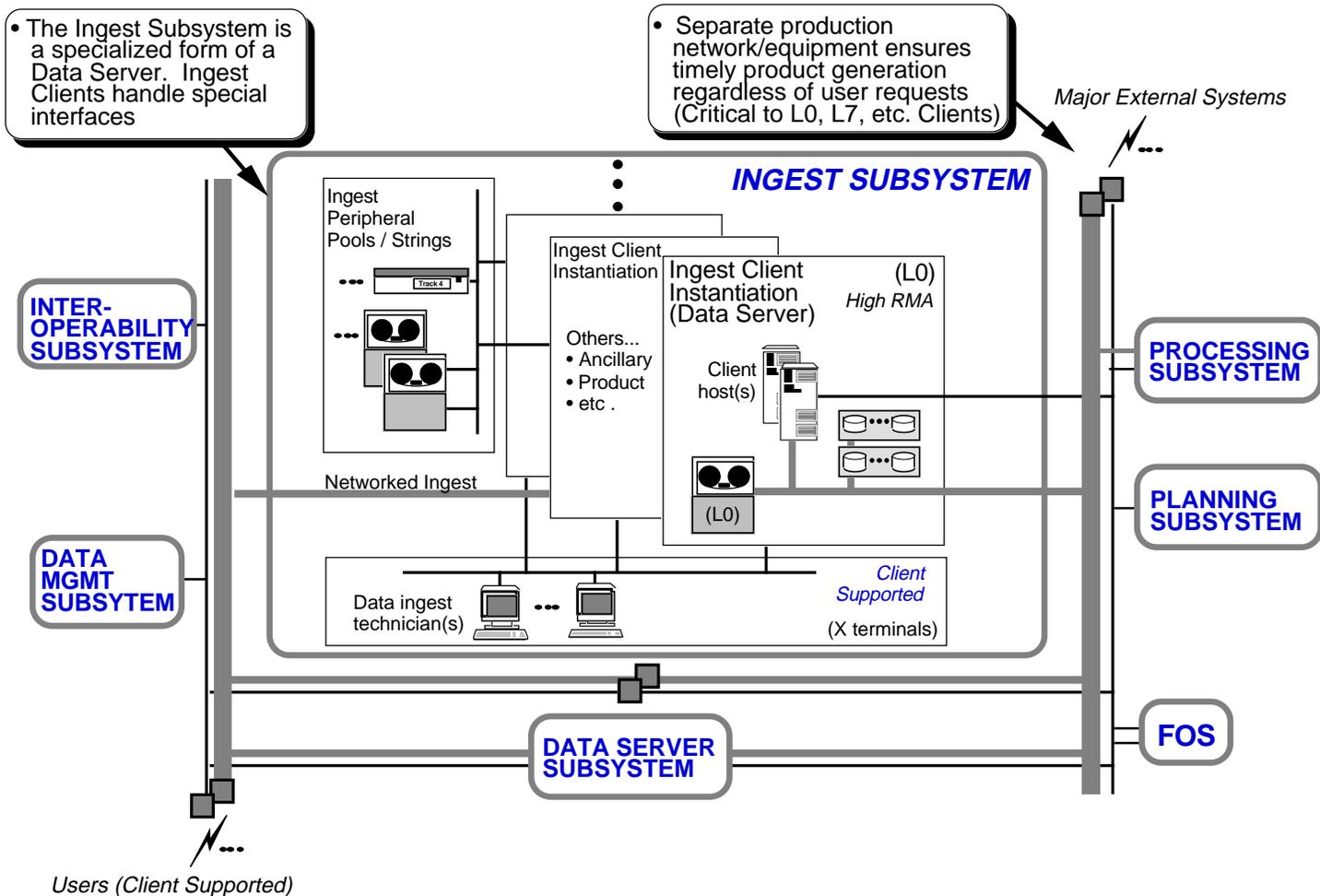


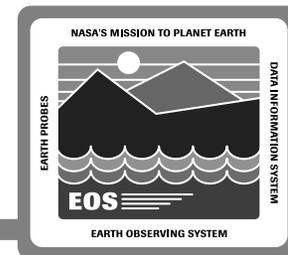
Data Server Scalability

| Site | PAM Robotics Increase (%) | PAM Drive Increase (%) | Working | Storage | Dist. | Staging | Data Server Scalability Strategies (Applied against "min" & "max" "pull" model loads and *2 volume size) (Release B) |
|-------|---------------------------|------------------------|-----------------------|-----------------------|----------------------------|--------------------------|--|
| | | | 1st Tier Increase (%) | 2nd Tier Increase (%) | Dist. Staging Increase (%) | Access Area Increase (%) | |
| LaRC | 200% (from 1 to 2) | 233% (from 3 to 7) | 230% | 200% | 276% | 276% | <ul style="list-style-type: none"> • Working Storage 2nd tier increase to handle larger production and volume loads • Working Storage & Distribution staging increases • Number of physical repositories increase (robotics) • Increase in user access requires more robotics (capacity and robotics not necessarily coupled) |
| NSIDC | 300% (from 1 to 3) | 300% | 288% | 0% | 288% | 288% | <ul style="list-style-type: none"> • Small site requirements still in envelope of 2nd tier capabilities • Working Storage & Distribution staging increases • Number of physical repositories increase (robotics) • Increase in user access requires more robotics (capacity and robotics not necessarily coupled) -- Small volume, but heavily accessed at small sites |



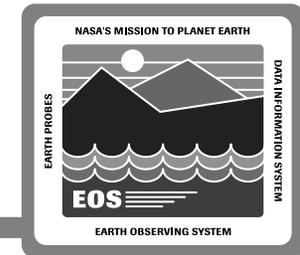
Data Ingest Subsystem





Data Ingest Issues / Features

| Major Subsystem Issue Description | Supporting Design Features |
|---|--|
| Ingest Subsystem high availability | <ul style="list-style-type: none"> • High RMA built into subsystem design • Network-attached devices as required • Parallel prime/backup hardware as required • RAID storage as required |
| Ingest Subsystem / EDOS buffering schemes , including the ability to ingest in “catch-up” mode (outage handling). | <ul style="list-style-type: none"> • "Rolling storage" (archive data repository) expands effective online storage • Transparent migration of ingested working data to data repositories sized to meet unique needs |
| Ingest Subsystem ability to accommodate unique interfaces and support data format translations and data transformation as needed. | <ul style="list-style-type: none"> • Unique Ingest Client for each data type (or similar data types). • Ingest parameters tuned for each interface and/or data type. |



Data Ingest Sizing

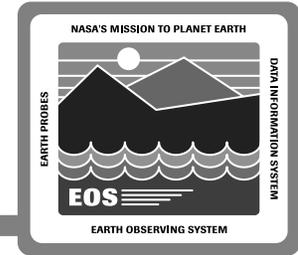
| Site | Ingest Hosts (# units) | WorkingStorage (# units / capacity per) | Number of Data Repository drives | Number of Data Repository Robotics |
|-------|------------------------|---|----------------------------------|------------------------------------|
| ASF | TBD | TBD | TBD | TBD |
| EDC | 3 | 4/96 | 3 | 1 (Large) |
| GSFC | 2 | 3/48 | 3 | 1 (Small) |
| JPL | 2 | 2/12 | 0 | 0 |
| LaRC | 2 | 3/48 | 3 | 1 (Small) |
| MSFC | 2 | 2/12 | 3 | 0 (sneaker net) |
| NSIDC | 2 | 2/12 | 0 | 0 |
| ORNL | 2 | 2/12 | 0 | 0 |

JPL, NSIDC, ORNL: In Release B, higher level products are brought into DS repositories. Ingest Clients handle necessary QA & data transformations as necessary.

ASF, NSIDC, JPL, ORNL: No repositories in Ingest due to lack of L0 ingest requirements at that site

Definitions

Robotics LARGE = 150 TB *Capacity* per
 Robotics SMALL = 25 TB *Capacity* per



Data Ingest Scalability Strategies

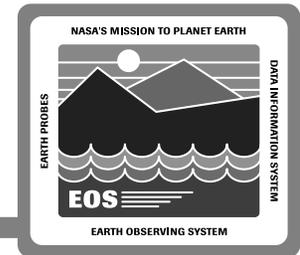
Increases in Ingest Volume

- **Ingest Data Server WS and Access layers scale in same ways as for any other Data Server**
- **High RMA must be maintained**

Increases in Long Term Storage Volume

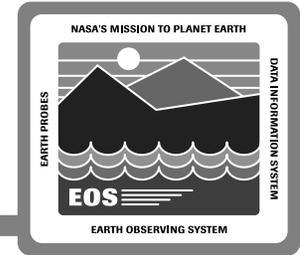
- **Ingest Data Server Data Repositories scale in same ways as for any other Data Server**
- **Deep archives would inherit Permanent Archive Management (PAM) data maintainability requirements**

Data Ingest Scalability



| Site | PAM Robotics Increase (%) | PAM Drive Increase (%) | Working Storage (%) | Ingest Scalability Strategies (Applied against *2 volume size) (Release B) |
|-------|---------------------------|------------------------|---------------------|---|
| LaRC | 0% | 0% | 200% | <ul style="list-style-type: none"> Working Storage staging increases due to volume increase Number of physical repositories: no change, within envelope (robotics & capacity) |
| NSIDC | n/a | n/a | n/a | <ul style="list-style-type: none"> n/a -- no L0 ingest, all physical ingest handled by Data Servers |

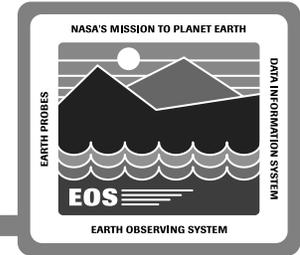
“Small” DAAC Topology: NSIDC



See Next Page



“Large” DAAC Topology: LaRC



See Next Page

