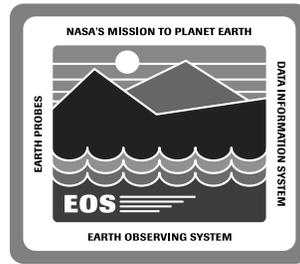


440-TP-006-001



# Production Topologies: A Trade-off Study Analysis for the ECS Project

Technical Paper

Technical Paper—Not intended for  
formal review or government approval.

April 1995

Prepared Under Contract NAS5-60000

## RESPONSIBLE ENGINEER

N. Prasad /s/	4/6/95
<hr/>	
Narayan S. Prasad, PDPS Scientist/Engineer EOSDIS Core System Project	Date

## SUBMITTED BY

Parag N. Ambardekar /s/	4/6/95
<hr/>	
Parag Ambardekar, PDPS Manager EOSDIS Core System Project	Date

Hughes Applied Information Systems  
Landover, Maryland

This page intentionally left blank.

# Abstract

---

Within the current processing architecture framework, chains of processing resources are pooled or dedicated to support Ad Hoc Working Group on Production (AHWGP) requirements. These equipment chains called "string" are composed of hardware laid out to perform processing, reprocessing, science software integration and test and backup processing, while minimizing impacts on communications and other staging infrastructure wherever possible. The concept of processing clusters has yielded a series of recommended physical processing topologies that can impact hardware requirements, overall performance, network capacity, staging storage, product generation throughput, etc.

The first phase of this trade-off analysis study examines the "One instrument's products per cluster" optimization alternative with data from the AHWGP for LaRC for epoch F. The ECS Systems Performance Model was used to dynamically simulate the processing. The remaining alternatives will be analyzed during the CDR phase. Recommendations will then be made for the most cost effective way of distributing processing to maximize throughput, minimize data movement, and provide and retain the flexibility to evolve with changing processing requirements.

**Keywords:** Topologies, cluster optimization alternatives, ECS System Performance Model, string, subnetwork, processing, dynamic analysis, static analysis

# Contents

---

## 1. Introduction

1.1	Trade Description .....	1-1
1.2	Scope.....	1-1
1.3	Organization .....	1-1
1.4	Acknowledgments.....	1-2
1.5	Review and Approval.....	1-2
1.6	Applicable and Reference Documents.....	1-2

## 2. Executive Summary

2.1	Major Analysis/Implementation Alternatives.....	2-1
2.2	Analysis Summary.....	2-2

## 3. Background

3.1	Background of Data Processing Subsystem.....	3-1
3.2	Background of ECS Systems Performance Model .....	3-1
3.3	AHWGP Data.....	3-3

## 4. Cluster Optimization Alternatives

4.1	Introduction .....	4-1
4.2	One Instrument's Products Per Cluster.....	4-2
	4.2.1 Physical View.....	4-2
4.3	One Instrument's Products Per Cluster Except for Selected Products Requiring Major Processing Resources .....	4-3
	4.3.1 Physical View.....	4-3
4.4	Multiple Instruments' Products on Any Cluster .....	4-4

4.4.1	Physical View.....	4-4
4.5	Any Instruments Products on Any Cluster That Can Support It.....	4-4
4.5.1	Physical View.....	4-4

## 5. Analysis

5.1	General Assumptions .....	5-1
5.2	Processing Requirements by Instrument at LaRC.....	5-1
5.2.1	CERES.....	5-1
5.2.2	MISR.....	5-4
5.2.3	MOPITT.....	5-5
5.3	Processing Requirements at LaRC .....	5-6
5.4	Factors for Cluster Optimization Alternatives .....	5-7
5.5	Analysis by Optimization Alternative.....	5-7
5.5.1	One Instrument's Products Per Cluster.....	5-7
5.5.2	One Instrument's Products Per Cluster Except for Selected Products Requiring Major Processing Resources.....	5-22
5.5.3	Multiple Instruments' Products Per Cluster.....	5-22
5.5.4	Any Instrument's Products on Any Cluster That Can Support It .....	5-23

## 6. Conclusions

### Abbreviations and Acronyms

### Figures

3.2-1.	Top-level Module of ECS Systems Performance Model.....	3-2
4.1-1.	Generic Star Topology of a Cluster .....	4-1
4.2-1.	Physical View of One Instrument's Products Per Cluster .....	4-2
4.3-1.	Physical View of One Instrument's Products Per Cluster Except For Selected Products Requiring Major Processing Resources.....	4-3
4.4-1.	Physical View of Multiple Instruments' Products On Any Cluster.....	4-4
4.5-1.	Physical View Of Any Instruments On Any Cluster That Can Support It.....	4-5

5.5-1. Schematic of Data Movement at CPU.....	5-11
5.5-2. CERES Processing Resource Usage.....	5-17
5.5-3. MISR Processing Resource Usage.....	5-19
5.5-4. MOPITT Processing Resource Usage.....	5-20

## Tables

5.2-1. CERES Requirements Summary.....	5-1
5.2-2. Summary of Definitions.....	5-3
5.2-3. MISR Requirements Summary.....	5-4
5.2-4. MOPITT Requirements Summary.....	5-5
5.3-1. Daily Average I/O and CPU Requirements at LaRC.....	5-6
5.5-1. CERES Daily Average Processor Requirements as a Function of Duty Cycle.....	5-8
5.5-2. MISR Daily Average Processor Requirements as a Function of Duty Cycle.....	5-9
5.5-3. MOPITT Daily Average Processor Requirements as a Function of Duty Cycle.....	5-10
5.5-4. CERES Daily Average Theoretical I/O Bandwidth at CPU and Processor <--> Data Handler Throughput.....	5-13
5.5-5. MISR Daily Average Theoretical I/O Bandwidth at CPU.....	5-14
5.5-6. MOPITT Daily Average Theoretical I/O Bandwidth at CPU and Processor <--> Data Handler Throughput.....	5-14
5.5-7. CERES Process Completion Times.....	5-15
5.5-8. CERES CPU and Staging Disk Capacity from ECS Systems Performance Model.....	5-17
5.5-9. MISR Process Completion Times.....	5-18
5.5-10. MISR CPU and Staging Disk Capacity from ECS Systems Performance Model.....	5-19
5.5-11. MOPITT Process Completion Times.....	5-20
5.5-12. MOPITT CPU and Staging Disk Capacity from ECS Systems Performance Model.....	5-21

# 1. Introduction

---

## 1.1 Trade Description

Within the current processing architecture framework, chains of processing resources are pooled or dedicated to support Ad Hoc Working Group on Production (AHWGP) requirements. These equipment chains previously called "string" within the SDPS SDS (also called cluster or subnetwork) are composed of hardware laid out to perform processing, reprocessing, science software integration and test, and backup processing, while minimizing impacts on communications and other staging infrastructure wherever possible. Each individual cluster may be designed to support the unique processing and reprocessing requirements of product generation tasks assigned to it. The concept of processing clusters has yielded a series of recommended physical (not logical) processing topologies that can impact hardware requirements, overall performance, network capacity, staging storage, product generation throughput, etc. This trade examines the pros and cons of distributing processing tasks from one or more instruments across one or more processing clusters. Recommendations will be made for the most cost effective way of distributing processing to maximize throughput, minimize data movement, and provide and retain the flexibility to evolve with changing processing requirements.

## 1.2 Scope

In the first phase of this trade study, a static and a preliminary dynamic analysis of AHWGP data will be performed for the third quarter of 1999 (Release B/C) time period. Given the number of instruments and the amount of data to be processed, the LaRC DAAC provides an ideal case for this study. The LaRC DAAC will process CERES, MISR and MOPITT instrument data. Because CERES and MISR have large processing requirements and I/O loads, and together with MOPITT have external data dependencies (e.g. MODIS products), they will provide insight into the architectural constraints, if any for the production topologies considered. The ECS System Performance Model will be used to dynamically simulate processing based on requirements provided by the AHWGP.

This paper is intended to set the stage for a more detailed dynamic analysis planned during the CDR phase. Recommendations will be made then.

## 1.3 Organization

This paper is organized as follows:

An executive summary provides an outline of major analysis, implementation alternatives, and preliminary results reported in this technical paper. Section 3 gives a background of the Data Processing Subsystem, the ECS System Performance Model and AHWGP data. The various cluster optimization alternatives are discussed in Section 4. Section 5 details the analysis performed based on the AHWGP data for all three LaRC instruments, namely CERES, MISR and MOPITT.

Preliminary results from the dynamic analysis using the System Performance Model is presented. Advantages and disadvantages of the various optimization alternatives are also listed. Section 6 draws conclusions based on the analysis. Acronyms used throughout the text are listed in Section 7.

## **1.4 Acknowledgments**

The assistance given by the ECS System Performance Modeling team (Bob Howard, Mike Theobald and Rajesh Dharia) is sincerely acknowledged. Eric Dodge deserves appreciation in providing ideas and valuable comments for this study.

## **1.5 Review and Approval**

Questions regarding technical information contained within this paper should be addressed to the following ECS and/or GSFC contacts:

- ECS Contacts  
Narayan Prasad, PDPS Scientist/Engineer  
(301)-925-0467  
nprasad@eos.hitc.com
- GSFC Contacts  
Steve Kempler, PDPS Manager  
(301)-286-7766  
steven.j.kempler@gsc.nasa.gov

Questions concerning distribution or control of this document should be addressed to:

Data Management Office  
The ECS Project Office  
Hughes Applied Information Systems  
1616 McCormick Dr.  
Landover, MD 20785

## **1.6 Applicable and Reference Documents**

1. Science Data Processing Segment (SDPS) Segment Design Specifications for the ECS Project, 305-CD-002-001.
2. Systems Performance Model for the ECS Project, 241-TP-001-001.
3. Investigator guide to estimating EOS data production. Bruce Barkstrom and members of the Ad Hoc Working Group on Production (unpublished).
4. Network Attached Storage Concepts and Industry Survey, 440-TP-009-001.
5. Trade-off Studies Analysis Data for the ECS Project, 211-CD-001-002.

## 2. Executive Summary

---

### 2.1 Major Analysis/Implementation Alternatives

In the first phase of this trade (this trade will continue into CDR), a static and a preliminary dynamic analyses of the AHWGP data is made for the Release B/C time period during the third quarter of 1999 (epoch F) at the LaRC DAAC. This time period is chosen because it is when maximum capacity requirements are exercised for all three instruments (CERES, MISR and MOPITT) scheduled to be processed at this site. Projected to be the largest DAAC with the most instruments, LaRC is currently expected to need 7230 MFLOPS (Millions of Floating Point Operations Per Second) with near maximum capacity for processing data from these instruments (these are raw numbers for Standard Processing only). The AHWGP data are also input into the ECS Systems Performance Model. Only Standard Processing is considered for this analysis. Combinations of processing scenarios with various processing topologies will be explored to serve as a guideline to determine optimized configurations of processing hardware. It should be noted that the calculations are based on the November 1994 AHWGP baseline. The AHWGP numbers may be refined if processing requirements change.

It is important to describe the concept of clusters from an operational viewpoint. Clusters are physical optimizations that do not prevent wholesale pooling for processing or reprocessing campaigns. The concept of processing clusters, current performance requirements, the resource pooling and dedication trade analysis at SDR has yielded different candidate cluster formations or cluster formation optimization alternatives which optimize different selection criteria (e.g. communications, staging, RMA (Reliability, Maintainability, Availability), ease of operations, management and control). It should be emphasized that these alternatives are fully configurable on a case-by-case basis, making it flexible to handle changing requirements or by release. The overall Planning and Data Processing architecture is built on the concepts of resource pooling regardless of the physical network layout. The different subnetwork/string/cluster formation alternatives are:

- o *One instrument's products per cluster* -- In this option, each instrument has a processing cluster, consisting of one or more compute servers dedicated to the production of its products.
- o *One instrument's products per cluster except for selected products requiring major processing resources* -- This option is identical to the first option except that at certain times when processing resources for certain products of that instrument exceeds the maximum allowable resources of that processing cluster, then processing of that product can shift to the cluster that has the resources to support it. Dynamic analysis for this option is deferred until CDR for the Systems Performance Model to be upgraded to handle this situation. Only a static analysis of the AHWGP data is performed.
- o *Multiple instruments' products on any cluster* -- This may apply to conditions whereby instruments with interdependent processing may be collocated. This situation does not

apply to the LaRC scenario. There is no interdependency among products from the three instruments. Also, the MODIS products that each instrument requires are different.

- o *Any instruments' products on any cluster that can support it* -- This option is a mix-and-match situation. The processing load will determine the cluster where a particular instrument's products will be processed. This option needs a full scale dynamic analysis. It will be explored after PDR when the Systems Performance Model is made sophisticated enough to handle this.

The allowable optimized cluster formation alternatives may:

- o be more than one at any given site. This may be the case when a site handles many instruments, and the processing requirements among instruments show large spread (e.g. MOPITT's processing requirements are small compared to CERES and MISR);
- o differ from one site to another because each DAAC handles different instruments;
- o differ with each release when significantly more complexities are introduced into the system and processing requirements increase.

## 2.2 Analysis Summary

Physical cluster optimization, on a site-by-site basis for release A, is not a major concern due to the small numbers and scale of the physical equipment currently envisioned for activation at that time. Release A implementations predicted for operations for LaRC and MSFC involve mid-performance (predicted) LANs and only two physical science processors within the SPRHW CI [1]. The GSFC configuration, which does not support processing operations, involves one or a small number of compute resources at a maximum. Thus, single physical subnetworks can be used (with the proper backup for RMA concerns) to couple the processing resources with primary ingest and Data Server resources, for example. The driver on selecting more than one subnetwork will be the actual throughput rates required, as opposed to operational or mission requirements which form the real basis of the implementation alternatives summarized earlier. It is expected that physical subnetwork optimization will be a larger issue for releases B and beyond. Therefore this analysis, which provides interim results for release-A, does focus ahead to release B and beyond.

The key recommendation is that multiple strings/cluster/subnetwork formation alternatives and selection criteria be allowed both between DAAC sites and within. One implementation alternative for all sites and all releases is not recommended. This will permit subnetworks of ECS resources to be tuned to meet the primary needs of the DAAC site, but will not disallow the view of the resources (through planning and production management) as a single processing pool or series of subpools.

The first option (one instrument's products per cluster) is a more natural way of doing data processing. Based on the requirements from the AHWGP, the three instruments can be each assigned to independent clusters. An obvious disadvantage of this set up is that the processing resources on a cluster may not be fully utilized, while a backlog can occur on another. As an example, a static analysis of the AHWGP data has yielded the following. MOPITT L1 and L2 processes are activated only once a day, while L3 processes are activated only weekly. With daily

average MFLOP requirements for MOPITT at less than 20 (based on raw numbers), it is a good candidate for sharing resources with MISR or CERES.

A static analysis of the AHWGP data has yielded a daily average MFLOP requirements for MISR for this time period to be 3450, with an "I/O bandwidth at CPU" of 18 MB/s. Since MISR requires large volumes of data to be staged (4 MB/s) and destaged (2.2 MB/s) (due to the large number of activations per day), it appears that distributing (according to optimization alternatives discussed earlier) MISR processing can increase network traffic, which in turn can degrade overall performance.

The CERES Data Processing Subsystems 4 (determine cloud properties, and top of the atmosphere and surface fluxes) and 5 (compute surface and atmospheric fluxes) take up more than 90% of the total CERES MFLOP requirements (1962 and 1781 MFLOPS for Subsystems 4 and 5, respectively). The daily average I/O bandwidth at CPU for Subsystems 4 and 5 are 5.4 MB/s and 0.6 MB/s, respectively. With relatively low I/O bandwidth at CPU, other CERES Subsystems (excluding 4 and 5) could share resources with MOPITT.

The ECS Systems Performance Model is used under the first configuration (one instrument's products per cluster). The volume of the staging disk and the optimal number of processors are analyzed for each instrument.

This page intentionally left blank.

## 3. Background

---

### 3.1 Background of Data Processing Subsystem

The Data Processing Subsystem (DPS) consists of three hardware CIs namely: 1) Science processing (SPRHW), 2) Algorithm Integration and Test (AITHW), and 3) Quality Assessment and Monitoring (AQAHW). It is responsible for managing, queuing and executing processes on a specified set of processing resources at each DAAC site. The science processing resources can be a chain of processing resources known as "clusters". They are self-contained processing resources based on a set series of alternatives for selection. They may also imply chains of processing, I/O and staging resources configured to deal with unique processing requirements to which they are allocated. This does not imply that the only use of that processing cluster, or a specific compute server on that cluster, is for only one specific class of instrument algorithms alone. It should be emphasized that cluster formations are configurable on a case-by-case basis, making it fully flexible to handle changing requirements by release. Clusters can be used for both processing and reprocessing. One or more data servers stage data on to the relevant working storage pool allocated to the processing clusters. Separate cluster resources are allocated for algorithm integration and test. This cluster called the Test and Backup cluster is configured with a like complement of processing, I/O and attached staging resources. Each processing cluster is supported by the Test and Backup cluster. The cluster topology provides a "fail soft" environment almost by its very nature.

### 3.2 Background of ECS Systems Performance Model

The ECS Systems Performance Model is a Block Oriented Network Simulation (BONeS) model [2] which is used in conjunction with the AHWGP data to simulate processing. Figure 3.2-1 contains the top-level module of BONeS with representation of ECS Subsystems. A brief description of the model components follow:

BONeS is a discrete-event simulation tool for analysis and design of communication networks and distributed processing systems. The components of a distributed processing system (including the networks) are represented by nodes. Nodes have resources associated with them which get allocated as events request them. Standard production of instrument data within DPS is simulated by the Processing module, in conjunction with Event Driven Scheduler and the Data Handler. The Data Handler is the model's representation of the Data Server design. It is responsible for storing and retrieving data from the permanent archive, for routing data to the requesting subsystems, and for managing tiered storage resources. The scheduler monitors the availability of data, requests data to be staged from the data Handler to Processing, routes newly created data to the appropriate data handler or processing pool, and initiates execution of a process when all required inputs are present. The Ingest module emulates behavior of the Ingest subsystem: acceptance of data from external systems and users, rolling storage of L0 instrument data, etc. are handled here. The Distribution module simulates network and media distribution of data to users.

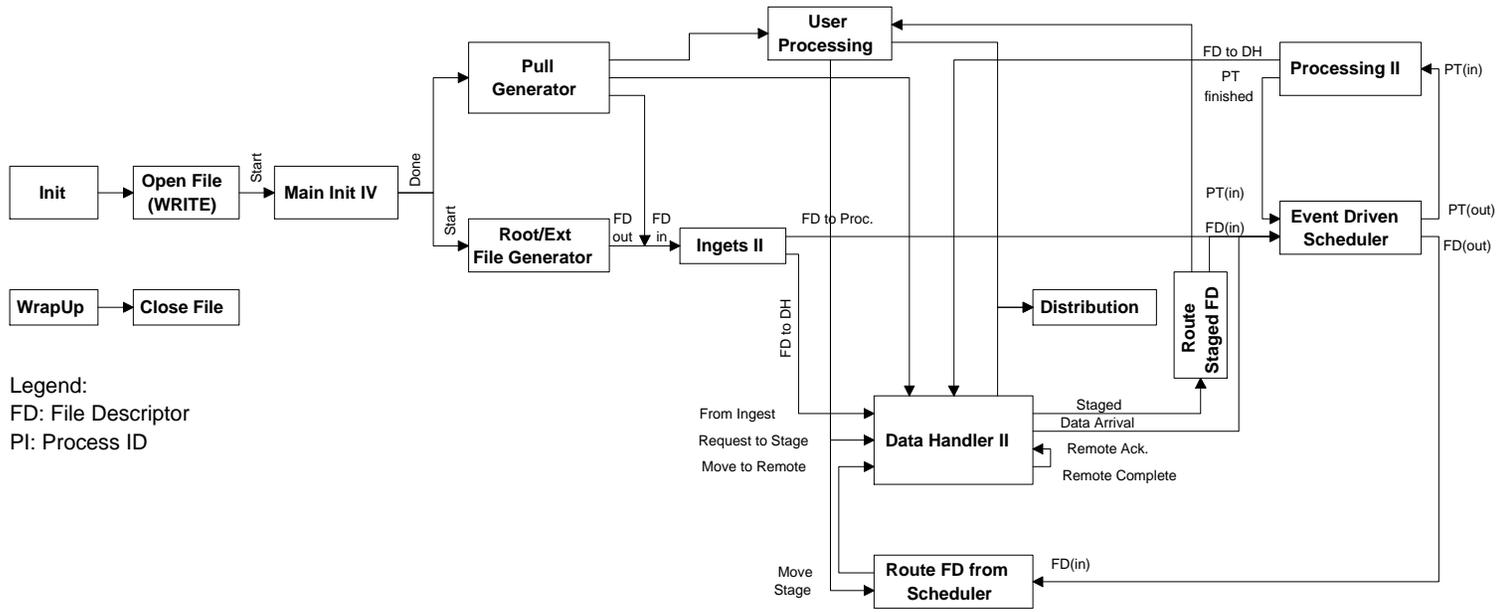


Figure 3.2-1 Top-level Module of ECS Systems Performance Model

During a simulation, the program collects data at selected points in the model network using a variety of probes. The model simulation is "resource constrained" in that the nodes in the model are specified to correspond to a particular system configuration. For example, the number of processors in a processing cluster may be constrained. The model will then determine for the constrained number of processors the time required to handle the data volume for normal operations. The performance of this configuration is measured by the simulation. The model is currently tested to verify operations within the entire suite of AHWGP data. The model will incrementally be made sophisticated to support an array of planned experiments and design tradeoffs.

### **3.3 AHWGP Data**

The Ad Hoc Working Group on Production [3] is represented by members of ASTER, CERES, LIS, MISR, MODIS, MOPITT and other instruments. The AHWGP was formed to produce a reliable estimate of the computer and network resources required to support ECS data production, and to provide ECS modelers and the Project information on data production plans. More specifically data from the AHWGP includes data products and their sizes, names of processes, number of activations and their activation scenarios, an estimate of the CPU processing capacity for each process, staging disk storage, number of file transfers and their sizes.

This page intentionally left blank.

# 4. Cluster Optimization Alternatives

## 4.1 Introduction

Clusters are physical optimizations that do not prevent wholesale pooling for processing or reprocessing campaigns. The concept of processing clusters, current performance requirements, the resource pooling and dedication trade analysis at SDR (unpublished) have yielded different candidate cluster formations or cluster formation optimization alternatives which optimize different selection criteria (e.g. communications, staging, RMA, ease of operations, management and control). It should be emphasized that these alternatives are configurable on a case-by-case basis, making it fully flexible to handle changing requirements or by release. The overall Planning and Data Processing architecture is built on the concepts of resource pooling regardless of the physical network layout. Figure 4.1-1 illustrates a generic star topology of a cluster formation. This topology is presented only as an example and does not imply that DAAC hardware will be configured this way. The DAAC topologies are driven by DAAC unique requirements (see Section A.1-A-3 in the DAAC Unique Appendices for operational sites [1]).

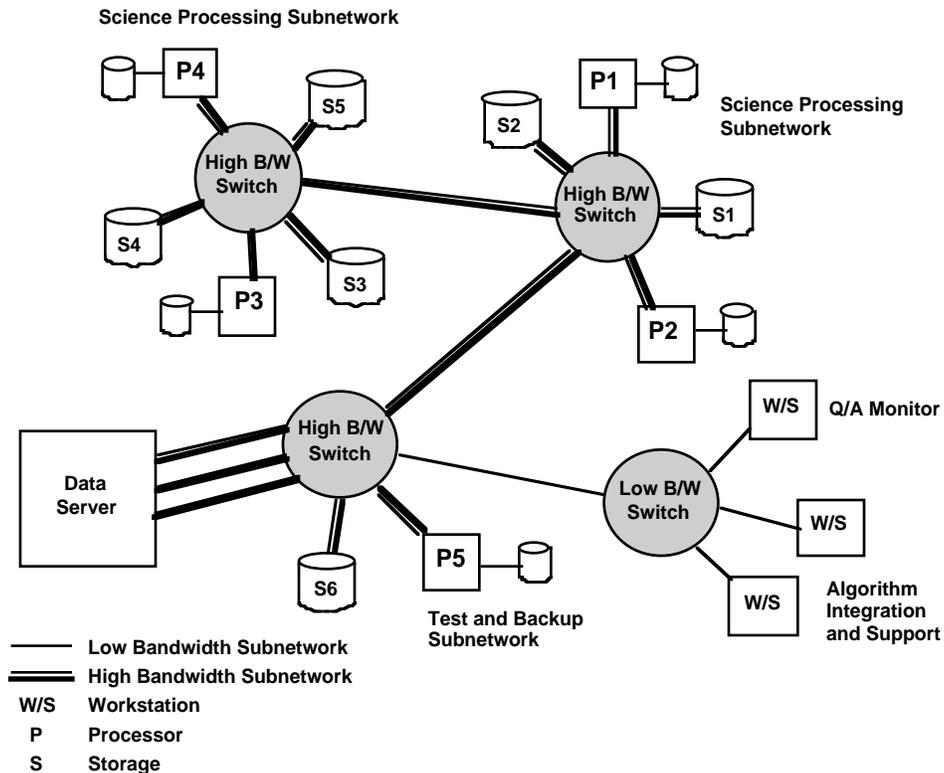


Figure 4.1-1. Generic Star Topology of a Cluster

The allowable optimized cluster formation alternatives may:

- o be more than one at any given site. This may be the case when a site handles many instruments, and the processing requirements among instruments show large spread (e.g. MOPITT's processing requirements are small compared to CERES and MISR);
- o differ from one site to another because each DAAC handles different instruments;
- o differ with each release when significantly more complexities are introduced into the system and processing requirements increase.

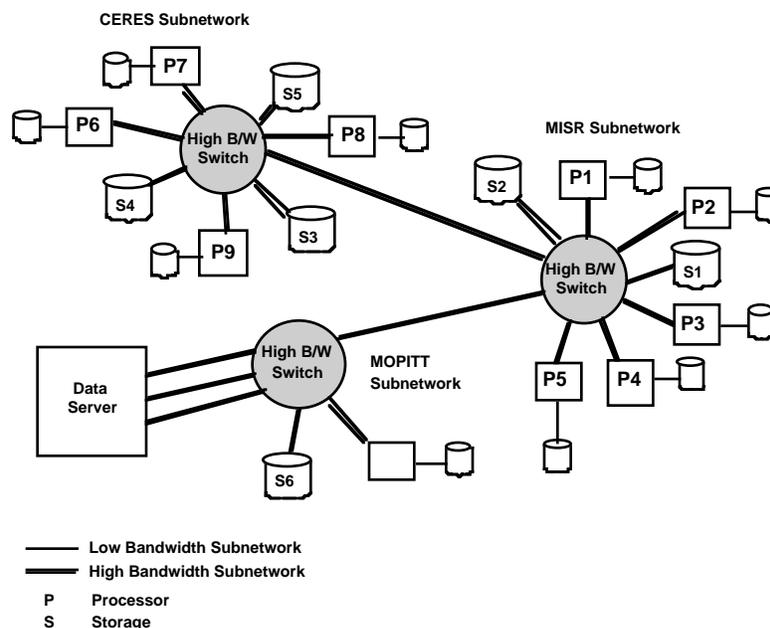
The different subnetwork/string/cluster formation alternatives are discussed in the following sections:

## 4.2 One Instrument's Products Per Cluster

In this option, each instrument has a processing cluster consisting of one or more compute servers dedicated to the production of its products.

### 4.2.1 Physical View

Physically each cluster is separate. They process data specific to an instrument. Instrument specific Product Generation Executives (PGEs) run on the processors comprising a cluster assigned to an instrument. Each cluster is self-contained and contains all the infrastructure necessary to process a particular instrument. Figure 4.2-1 illustrates the physical view of the topology.



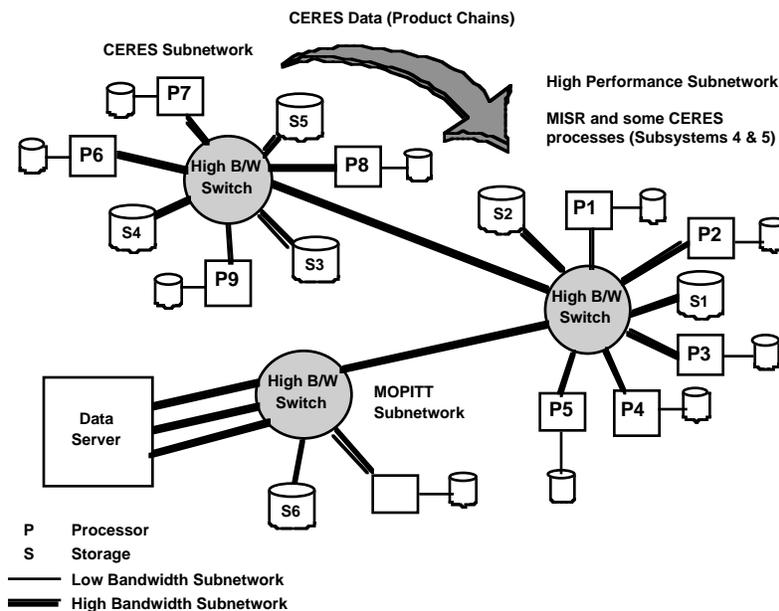
**Figure 4.2-1. Physical View of One Instrument's Products Per Cluster**

### 4.3 One Instrument's Products Per Cluster Except for Selected Products Requiring Major Processing Resources

This option is identical to the first option except that at certain times when processing resources for certain products of that instrument exceed the maximum allowable resources of that processing cluster, then processing of that product can shift to a cluster that has the available resources to support it.

#### 4.3.1 Physical View

A good example of this situation is pooling resources for MISR and CERES Subsystems 4 and 5 (largest subsystems in terms of processing requirements). Figure 4.3-1 shows a high performance subnetwork consisting of high performance machines exclusively for MISR and CERES Subsystems 4 and 5. This subnetwork shares two instruments. Since CERES Subsystems 4 and 5 require data generated by other CERES subsystems (product chains) on the CERES subnetwork, dependent CERES data need to be moved to the high performance subnetwork. Product chains are the concept of including dependent production of data products at higher levels. For example, it is the production of L1 data from L0 data using a L1 algorithm, L2 from L1 and so on. The production of some data products is dependent on other previous level data products and ancillary data products. It should be noted that although the subnetworks are physically separate, there is no logical separation.



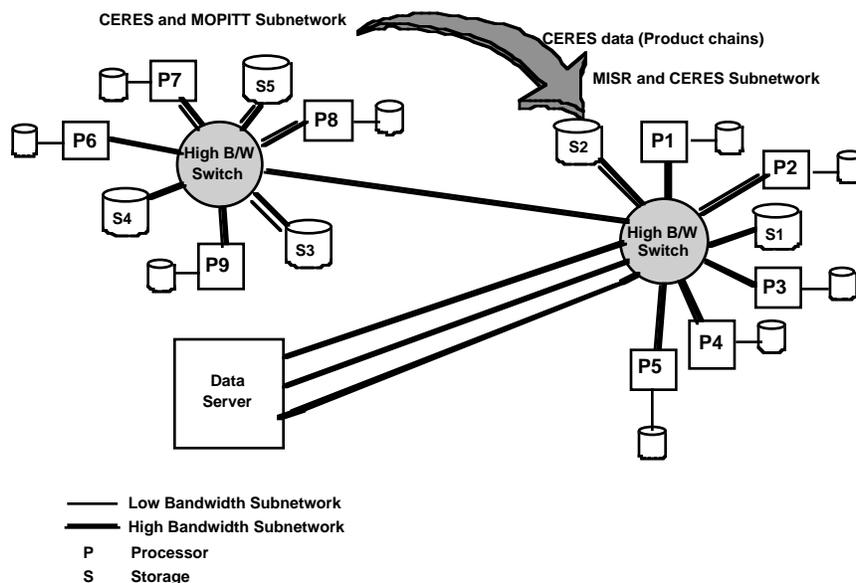
**Figure 4.3-1. Physical View of One Instrument's Products Per Cluster Except For Selected Products Requiring Major Processing Resources**

## 4.4 Multiple Instruments' Products on Any Cluster

This may apply to conditions whereby instruments with interdependent processing may be collocated. This situation does not apply for the LaRC scenario. There is no interdependency among products from the three instruments. Also, the MODIS products that each instrument uses are different.

### 4.4.1 Physical View

In the topology presented in Figure 4.4-1, each science processing network has sufficient resources to support multiple instruments. Data need to be moved from one subnetwork to another if there are product chain dependencies. Again, there is no logical separation of the subnetworks. They are only physically separated.



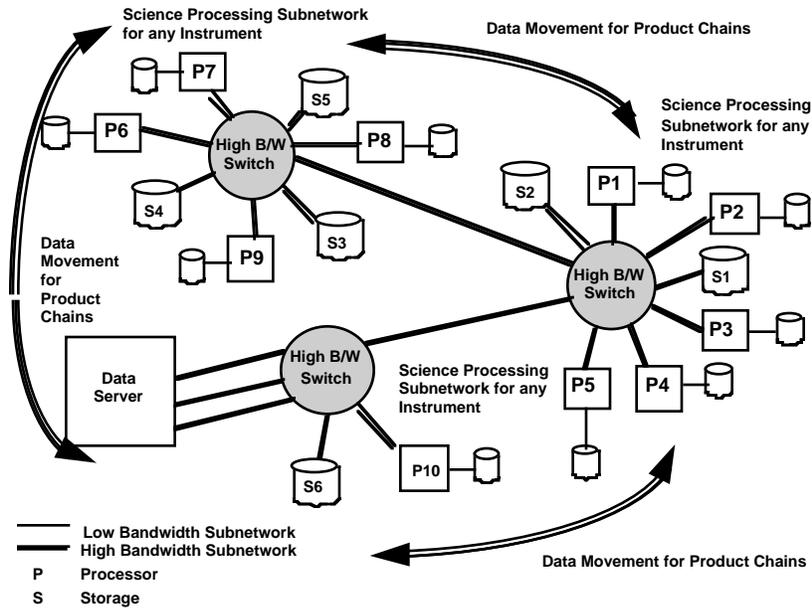
**Figure 4.4-1. Physical View of Multiple Instruments' Products On Any Cluster**

## 4.5 Any Instruments Products on Any Cluster That Can Support It

This option is a mix-and-match situation. The processing load will determine the cluster where a particular instrument's products will be processed.

### 4.5.1 Physical View

Figure 4.5-1 shows a topology where subnetworks may have resources to support any instrument.



**Figure 4.5-1. Physical View Of Any Instruments On Any Cluster That Can Support It**

This page intentionally left blank.

# 5. Analysis

---

## 5.1 General Assumptions

The following assumptions are made for the analysis of AHWGP data:

- The AHWGP (November 1994 baseline) data are representative of the kind of processing to be supported. As science algorithms are developed and launch dates near, these data are subject to change.
- Only generation of standard products is considered. Reprocessing will be considered as more AHWGP data become available, and the Systems Performance Model is made more sophisticated.
- A 24-hour time period is assumed.
- All estimates for computational load are based on Millions of Floating Point Operations (MFPOs). No distinction is made between floating point operations and non-floating point operations. They are two entirely different machine attributes that can vary within an architecture.
- Only raw numbers from the AHWGP are considered for this study. Input data are not scaled to represent the increase in processing requirements required for hardware selection.
- I/O control delay is not accounted for in static computations of processing times.

## 5.2 Processing Requirements by Instrument at LaRC

### 5.2.1 CERES

CERES data processing is organized into ten subsystems. These Subsystems are a logical collection of algorithms which together convert input data products into output data products. Table 5.2-1 lists CERES requirements per process with Table 5.2-2 providing definitions to the terms used in Table 5.2-1. Table 5.2-1 and following tables that represent the requirements summary are "rolled up" based on the numbers given by the AHWGP. It includes multiple instances of similar files. It also takes into account fractions of files read. As an example, for CERES there can be 240 instances of a file per activation. Subsystems 4 and 5 require the most data volumes at initiation and completion. They also have a substantial Millions of Floating Point Operations (MFPOs). In later sections we will see how these data volumes have implications for staging disk capacity.

**Table 5.2-1. CERES Requirements Summary (1 of 2)**

Process	Volume at Initiation (MB)	Staging I/O (MB)	Volume at Completion (MB)	Destaging I/O (MB)	I/O Reqts. (MB)	CPU Reqts. (MFPOs)	No. Input Files	No. Output Files	Activations (per day)
CERES 1aA	138	87	852	714	852	20,790	4	25	1
CERES 1aT	138	87	852	714	852	20,790	4	25	1
CERES 1bA	138	87	852	714	852	20,790	4	25	1
CERES 2aA	138	87	852	714	852	3,780	4	2	0.93
CERES 2aT	375	324	706	331	596	3,780	4	2	0.6
CERES 3aA	164	114	836	672	836	47,250	3	3	0.03
CERES 3aT	164	114	836	672	836	47,250	3	3	0.03
CERES 3bTA	278	227	949	672	949	94,500	4	3	0.03
CERES 4aF	348	206	593	245	505	34,020	8	2	24.8
CERES 4bA1F	2,768	2,626	3,013	245	3,013	3,402,000	8	2	24.8
CERES 4bA2F	2,768	2,626	3,013	245	3,013	3,402,000	8	2	24.8
CERES 5aF	205	154	426	221	426	2,672,460	3	1	24.8
CERES 5aV	205	154	426	221	426	2,672,460	3	1	4
CERES 5cAF	205	154	426	221	426	2,672,460	3	1	24.8
CERES 5cAV	205	154	426	221	426	2,672,460	3	1	4
CERES 6aA	271	221	275	4	275	4,914	3	1	24.8
CERES 6aT	271	221	275	4	275	4,914	3	1	24.8
CERES 7aA	2,168	2,118	3,484	1,316	3,484	680,400	330	40	0.2
CERES 7aT	2,168	2,118	3,484	1,316	3,484	680,400	330	40	0.2
CERES 7c	2,874	2,823	4,190	1,316	4,190	1,360,800	498	40	0.03
CERES 8aA	8,210	8,159	8,576	366	8,576	226,800	250	2	0.03
CERES 8aT	8,210	8,159	8,576	366	8,576	226,800	250	2	0.03
CERES 8c	8,210	8,159	8,576	366	8,576	453,600	250	2	0.03
CERES 9aAF	205	154	207	2	207	4,914	3	1	24.8
CERES 9aTF	205	154	207	2	207	4,914	3	1	24.8
CERES 10aA	10,317	10,367	10,931	564	10,931	245,700	1738	1	0.03
CERES 10aT	10,317	10,367	10,931	564	10,931	245,700	1738	1	0.03
CERES 10bTA	11,880	11,880	12,444	564	12,444	491,400	2482	1	0.03
CERES 11a	91	91	182	0	182	37,800	1	1	0.1
CERES 12aF	82	32	334	252	334	37,800	10	24	1
Total	73,716	72,224	87,730	13,824	87,532	22,493,646	7,955	256	237.5

**Table 5.2-1. CERES Requirements Summary (2 of 2)**

Process	Volume Staged (MB/day)	Volume Destaged (MB/day)	CPU requirements (MFPOs)/day
CERES 1aA	87	714	20,790
CERES 1aT	87	714	20,790
CERES 1bA	87	714	20,790
CERES 2aA	81	664	3,515
CERES 2aT	194	199	2,268
CERES 3aA	3	20	1,418
CERES 3aT	3	20	1,418
CERES 3bTA	7	20	2,835
CERES 4aF	5,109	6,076	843,696
CERES 4bA1F	65,125	6,076	84,369,600
CERES 4bA2F	65,125	6,076	84,369,600
CERES 5aF	3,819	5,481	66,277,008
CERES 5aV	616	884	10,689,840
CERES 5cAF	3,819	5,481	66,277,008
CERES 5cAV	616	884	10,689,840
CERES 6aA	5,481	99	121,867
CERES 6aT	5,481	99	121,867
CERES 7aA	424	263	136,080
CERES 7aT	424	263	136,080
CERES 7c	85	39	40,824
CERES 8aA	245	11	6,804
CERES 8aT	245	11	6,804
CERES 8c	245	11	13,608
CERES 9aAF	3,819	50	121,867
CERES 9aTF	3,819	50	121,867
CERES 10aA	311	17	7,371
CERES 10aT	311	17	7,371
CERES 10bTA	356	17	14,742
CERES 11a	9	0	3,780
CERES 12aF	32	252	37,800
Total	166,064	35,222	324,489,148

**Table 5.2-2. Summary of Definitions**

<b>Term</b>	<b>Definition</b>
Volume at initiation	Data volume expected at the start of a process initiation prior to each activation
Volume at completion	Data volume expected at the completion of a process after each activation. It is equal to volume at initiation + sum of all output file sizes for each activation. It includes temporary files.
Staging I/O	Volume staged from the archive for each activation of a process
Destaging I/O	Volume destaged to the archive after completion of a process (for each activation).
I/O requirements	Read and write operations from the staging disk per process for each activation.
CPU requirements	Millions of floating point operations per process per activation. Note that there is no time involved.
Number of input files	Number of input files each process requires per activation. This includes temporary files, input files from other instruments, or other lower level product files from the same instrument.
Number of output files	Number of output files each process produces per activation. This includes temporary files. Some output files may be input to other higher level processes.
Number of activations per day	The number of times the process is activated per day. If the number of activations is a fraction, then it is not activated every day. The process may be activated once every week. To get the number of times it is activated in a month, multiply the number of daily activations by 30.
Volume staged per day	The volume staged for each process times the number of activations of the process per day. This gives a daily average of the volume staged to a process.
Volume destaged per day	The volume destaged for each process times the number of activations of the process per day. This gives a daily average of the volume destaged after completion of a process.
CPU requirements per day	This is a daily average CPU requirements per process. It is obtained by multiplying CPU requirements for each process by the number of activations per day.

### 5.2.2 MISR

The MISR processing requirements summary is listed in Table 5.2-3 (the terms in each column are defined in Table 5.2-2). MISR data will be processed in units of one orbit. This translates to approximately 14.56 activations per day for all MISR processes. There will be four production software subsystems, one each for the products at Levels 1A, 1b, 2-T/C, and 2-A/S. Each of these will be capable of being operated individually, or as a combined unit that maximizes resources and throughput. Processing will not commence until the Planning Subsystem determines that all the data dependencies are satisfied. Operational data from external resources, e.g., meteorological data

from NOAA or instrument data from MODIS will be preprocessed by a separate element of the respective MISR software subsystems to prepare this data for use. The MISR team has not made any decisions yet on how the processing might be divided into individual processes. The four subsystems could be treated as a single entity or as separate entities for the purpose of resource allocation.

**Table 5.2-3. MISR Requirements Summary (1 of 2)**

Process	Volume at Initiation (MB)	Staging I/O (MB)	Volume at Completion (MB)	Destaging I/O (MB)	I/O Reqs. (MB)	CPU Reqs. (MFPOs)	No. Input Files	No. Output Files	Activations (per day)
MISP1A	3,167	3,167	7,451	4,284	7,451	237,000	23	76	14.5
MISP1B1	3,816	3,816	7,745	3,929	7,745	178,000	39	38	14.5
MISP1B2I	7,624	7,624	11,989	4,365	11,989	8,557,000	59	22	14.5
MISP2ASI	4,958	4,958	5,439	482	5,439	6,990,000	40	4	14.5
MISP2TCI	4,425	4,425	4,700	276	4,700	4,652,000	36	2	14.5
Total	23,990	23,990	37,324	13,336	37,324	20,614,000	197	142	72.5

**Table 5.2-3. MISR Requirements Summary (2 of 2)**

Process	Volume Staged (MB/day)	Volume Destaged (MB/day)	CPU requirements (MFPOs)/day
MISP1A	45,922	62,118	3,436,500
MISP1B1	55,332	56,971	2,581,000
MISP1B2I	110,548	63,293	124,076,500
MISP2ASI	71,891	6,989	101,355,000
MISP2TCI	64,163	4,002	67,454,000
Total	347,855	193,372	298,903,000

### 5.2.3 MOPITT

The scientific goals for MOPITT depend upon long term, homogeneous, global data products rather than the quick turn around of observations made at a particular geographical location. Therefore, the MOPITT Standard Product Generation algorithms will be designed in such a way that each level of processing is autonomous. Each level is dependent on the existence of its preceding level and upon the existence of certain ancillary data unique to a level. Processors for the various levels may be run in sequence or at different times assuming that the necessary dependencies are met. Table 5.2-4 lists the summary of requirements for generating MOPITT standard products. The terms in each column of Table 5.2-4 are defined in Table 5.2-2. It is

expected that L1 (Calibrated, Earth Located Radiance) and L2 (Retrieved Geophysical Parameters) data products will be produced on a daily basis (see activations per day column in Table 5.2-4). The minimum volume of data to be handled for each run will be data accumulated within the corresponding day. The L3 (Global, Gridded, Geophysical Parameters) products will be produced on a weekly basis.

**Table 5.2-4. MOPITT Requirements Summary (1 of 2)**

Process	Volume at Initiation (MB)	Staging I/O (MB)	Volume at Completion (MB)	Destaging I/O (MB)	I/O Reqs. (MB)	CPU Reqs. (MFPOs)	No. Input Files	No. Output Files	Activations (per day)
MOPL1	257	255	614	356	614	16,800	3	3	1
MOPL1Qi-D	356	356	366	10	366	900	2	1	1
MOPL2-E	405	344	589	185	589	1,502,250	12	3	1
MOPL2Qi-D	175	175	185	10	185	1,350	2	1	1
MOPL3	75	75	163	89	163	28,290	1	3	0.14
MOPL3Qi-F	79	79	89	10	89	900	2	1	0.14
Total	1,347	1,284	2,006	660	2,006	1,550,490	22	12	4.28

**Table 5.2-4. MOPITT Requirements Summary (2 of 2)**

Process	Volume Staged (MB/day)	Volume Destaged (MB/day)	CPU requirements (MFPOs)/day
MOPL1	255	356	16,800
MOPL1Qi-D	356	10	900
MOPL2-E	344	185	1,502,250
MOPL2Qi-D	175	10	1,350
MOPL3	11	12	3,961
MOPL3Qi-F	11	1	126
Total	1,152	575	1,525,387

### 5.3 Processing Requirements at LaRC

The LaRC DAAC may contain three processing clusters with each one corresponding to one instrument. The actual configuration may vary. MODIS products from EDC and GSFC are used as ancillary inputs. Table 5.3-1 lists the daily average I/O and CPU requirements at the LaRC DAAC at epoch F. The I/O requirements/day for CERES is obtained by multiplying the I/O requirement for each process by the number of activations per day summed over all processes. The CPU requirement/day for CERES is obtained by multiplying the CPU requirement for each

process by the number of activations per day summed over all processes. Similarly, the daily requirements for MISR and MOPITT are determined.

**Table 5.3-1. Daily Average I/O and CPU Requirements at LaRC**

Instrument	I/O Operations (MB/day)	CPU (MFPOs/day)
CERES	217,870	324,000.000
MISR	541,198	299,000,000
MOPITT	1,789	1,530,000

## 5.4 Factors for Cluster Optimization Alternatives

The choice of optimization alternative is based on the following factors.

- Processing requirements (MFLOPs) of various instruments;
- Product chains - production of data products at the next higher level. For example, it is the production of L1 data from L0 data using L1 algorithm, L2 from L1 and so on. The production of some data products is dependent on other previous level data products and ancillary data products;
- Special hardware or software requirements for each instrument;
- Process activations and estimated length of process runs;
- Anticipated growth in individual instrument requirements;
- Interdependency among multiple instruments (one instruments products is input to another instrument).

The following sections perform static and a preliminary dynamic analyses of each optimization alternative.

## 5.5 Analysis by Optimization Alternative

### 5.5.1 One Instrument's Products Per Cluster

Processing each instrument on a separate cluster is a natural extension of the processing clusters topology. Unique algorithm requirements may dictate the selection of particular compute resources that offer the best solutions.

### 5.5.1.1 Static Analysis

#### 5.5.1.1.1 Estimating Number of Processors and Staging Disk Capacity

A host attached disk is assumed for the calculations. The theoretical staging disk capacity can be estimated based on the "duty cycle" which can be defined as the amount of MFLOPs to be performed by a processor to the number of MFLOPs which a processor is capable. Processor performance is usually rated by vendors as peak MFLOPs. An efficiency factor ( $\eta = 0.25$ ) is used to adjust the vendor provided peak MFLOPs.

$$Duty\ cycle = MFPOs/process\ per\ day / (N \times MFLOPs/processor \times \eta \times 86400)..... (1)$$

where *MFPOs/process per day* : Millions of Floating Point Operations performed by the process per day,

*N*: Number of processors

*MFLOPs/processor* : Peak processor rating per processor.

A duty cycle of unity indicates a fully loaded processor.

The staging disk capacity can be written as:

$$Staging\ disk\ capacity\ (per\ day) = Vc \times t_{cp} .....(2)$$

where *Vc* : Data volume at completion (MB);

*t<sub>cp</sub>* : Estimated time of completion of a process.

#### 5.5.1.1.1.1 CERES

The relationship between duty cycle and the number of processors (for different peak MFLOP ratings) is illustrated in Table 5.5-1 for CERES based on total CPU requirements (see Table 5.3-1 converted to per sec). At 100-MFLOP peak rating, we need 150.2 processors at duty cycle 1 (ideally) to satisfy CERES CPU requirements. To be more realistic, this is equivalent to 187.7 processors at 80% duty cycle. Similarly, 50.1 processors with 300-MFLOPs peak rating at duty cycle 1 (ideal case) is equivalent to 62.5 processors at 80% duty cycle. With 1000-MFLOPs peak processors, we need only 3.8 processors (the difference of 18.8 and 15.0) to allow a 20% slack in duty cycle.

**Table 5.5-1. CERES Daily Average Processor Requirements as a Function of Duty Cycle**

Duty cycle	Number of 100-MFLOP processors	Number of 300-MFLOP processors	Number of 1000-MFLOP processors
1	150.2	50.1	15.0
0.95	158.1	52.7	15.8
0.90	166.1	55.6	16.7
0.85	176.7	58.9	17.6
0.80	187.7	62.5	18.8

If the duty cycle of a group of processors is held constant at unity, from equation 1 we see that by increasing the number of processors we can process more floating point operations per day. The capacity to process more floating point operations per day is related to throughput which in turn has implications to the staging disk capacity. If throughput can be increased, we can reduce the staging disk capacity. Let us do some theoretical calculations based on number of processors and determine staging disk capacity given in equation 2.

For a duty cycle of unity, to compute the daily average staging disk capacity we multiply the volume at completion for each CERES process by the number of activations per day (in Table 5.2-1 Part 2) and sum over all processes. The daily average staging disk capacity determined by this method yields 220,119 MB for CERES. This staging disk capacity is needed when 150.2 100-MFLOP (peak) processors are used at duty cycle unity. The theoretical estimate of staging disk capacity also indicates that 50.1 300-MFLOPs (peak) and 15.0 1000-MFLOPs (peak) processors are equivalent. This is true only in the static sense because dynamically, processes on a 1000-MFLOPs processor complete faster than on 100- or 300-MFLOPs processors, thereby requiring a smaller staging disk capacity. The static estimate of staging disk capacity may be an overestimate because it assumes that all processes start at the same time without accounting for disk volumes cleared by processes that end at various times. Nonetheless, it gives an upper bound for validating the dynamic model.

It is possible to determine a relationship between the number of processors and staging disk volume. From Table 5.3-1, based on the number of CPU requirements per day for CERES, we can calculate the time required for completion of all processes. If we use 60 300-MFLOP processors (at duty cycle 1) instead of 50.1, we can increase the processing capacity and reduce the time it takes for all CERES processes to complete. Therefore, the time of completion can be written as:

$$\begin{aligned}
 \text{Time of completion} &= \text{MFPOs per day} / (N \times \text{MFLOPs}_{\text{processor}} \times \eta) \dots\dots\dots(3); \\
 &= 3.24 \times 10^8 / (60.0 \times 300.0 \times 0.25 \times 3600 \times 24) \\
 &= 0.833 \text{ days}
 \end{aligned}$$

Similarly, substituting the appropriate numbers into equation (2):

$$\begin{aligned}
 \text{Staging disk capacity} &= 220,119 \text{ MB per day} \times 0.833 \text{ days} \\
 &= 183,432 \text{ MB} \equiv 183 \text{ GB}
 \end{aligned}$$

A savings of 16.7% in staging disk capacity can be obtained by adding 10 more 300-MFLOP processors and operating ideally at duty cycle 1. For more realistic duty cycles (less than 1), the staging disk capacity can be calculated similarly.

### 5.5.1.1.1.2 MISR

The relationship between duty cycle and number of processors (for different peak MFLOP ratings) is illustrated in Table 5.5-2 for MISR based on CPU requirements shown in Table 5.2-3. Ideally,

at 100-MFLOP peak rating, we need 138.4 fully dedicated processors (at duty cycle=1) to satisfy MISR CPU requirements. To be more realistic, this is equivalent to 173.0 processors at 80% duty cycle. Similarly, 46.1 processors with 300 MFLOPs peak rating at duty cycle = 1 (ideal case) is equivalent to 57.6 processors at 80% duty cycle. With 1000 MFLOPs peak processors, we need only 3.5 processors (the difference of 17.3 and 13.8) to allow a 20% slack in duty cycle.

**Table 5.5-2. MISR Daily Average Processor Requirements as a Function of Duty Cycle**

Duty cycle	Number of 100-MFLOP processors	Number of 300-MFLOP processors	Number of 1000-MFLOP processors
1	138.4	46.1	13.8
0.95	145.7	48.5	14.5
0.90	153.8	51.2	15.4
0.85	162.8	54.2	16.3
0.80	173.0	57.6	17.3

For a duty cycle of unity, to compute the daily average staging disk capacity we multiply the volume at completion for each MISR process by the number of activations per day (see Table 5.2 3) and sum over all processes. The daily average staging disk capacity determined by this method yields 541,198 MB for MISR. This staging disk capacity is needed when 138.4 100- MFLOP (peak) processors are used at duty cycle unity. The theoretical estimate of staging disk capacity also indicates that 46.1 300-MFLOPs (peak) and 13.8 1000-MFLOPs (peak) processors are equivalent. This is true only in the static sense because dynamically, processes on a 1000-MFLOPs processor complete faster than on 100- or 300-MFLOPs processors, thereby requiring a smaller staging disk capacity.

From Table 5.3-1, based on the number of CPU requirements per day for MISR, we can calculate the time required for completion of all processes. If we use 20 1000-MFLOP processors (at duty cycle 1) instead of 13.8, we can increase the processing capacity and reduce the time it takes for all MISR processes to complete. Substituting the appropriate numbers in equation (3) and (2), respectively:

$$\begin{aligned}
 \text{Time of completion} &= 2.99 \times 10^8 / (20.0 \times 1000.0 \times 0.25 \times 3600 \times 24) \\
 &= 0.692 \text{ days} \\
 \text{Staging disk capacity} &= 541,198 \text{ MB per day} \times 0.692 \text{ day} \\
 &= 374,579 \text{ MB} \cong 375 \text{ GB}
 \end{aligned}$$

A savings of 31% in staging disk capacity can be attained by adding 6 more 1000-MFLOP processors and operating ideally at duty cycle 1. For more realistic duty cycles, the staging disk capacity can be calculated similarly.

### 5.5.1.1.1.3 MOPITT

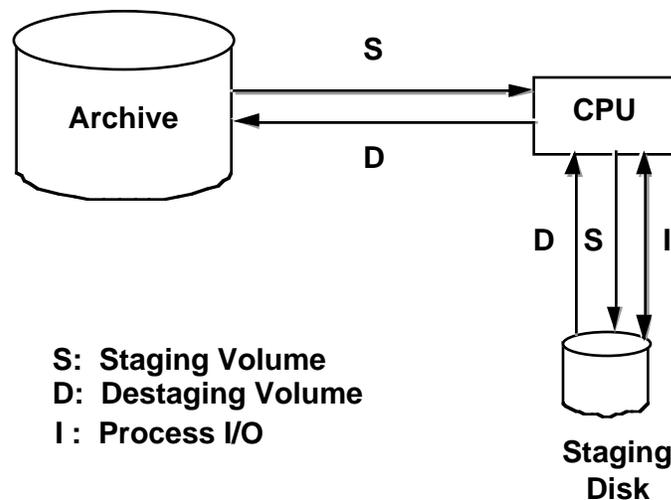
Table 5.5-3 gives MOPITT daily average processor requirements as a function of processor duty cycle based on CPU requirements given in Table 5.2-4. For MOPITT, a 20% slack in duty cycle can be accommodated by a single 100-MFLOPs peak processor.

**Table 5.5-3. MOPITT Daily Average Processor Requirements as a Function of Duty Cycle**

Duty cycle	Number of 100 - MFLOP processors
1	0.70
0.95	0.74
0.90	0.77
0.85	0.82
0.80	0.87

For a duty cycle of unity, to compute the daily average staging disk capacity, we multiply the volume at completion for each MOPITT process by the number of activations per day (in Table 5.2-4) and sum over all processes. The daily average staging disk capacity determined by this method yields 1,789 MB for MISR. This staging disk capacity is needed when 0.7 100-MFLOP (peak) processors are used at duty cycle unity. MOPITT staging disk volumes are not of major concern. However, we should determine ways where we can let other instruments share disk resources with MOPITT to reduce cost in buying a separate disk.

### 5.5.1.1.2 Static Estimates of I/O Bandwidth at CPU and Processor <--> Data Handler Throughput



**Figure 5.5-1. Schematic of Data Movement at CPU**

The I/O bandwidth at CPU gives a fairly good static estimate of the amount of data movement coordinated by a CPU. Figure 5.5-1 is a schematic depicting data movement among the archive, CPU and the staging disk. A host attached staging disk is assumed. Data movement can be significantly different for Network Attached Storage [4] [5]. When the Planning Subsystem determines that all dependencies are satisfied, it sends a Data Processing Request to the Data Processing Subsystem. Data are then staged to the staging disk by the processing CPU. After processing is completed, the output files that are to be archived are destaged from the staging disk. During processing the CPU coordinates the read/write operations of the process. Therefore, it is possible to theoretically estimate a daily average I/O bandwidth at CPU for each process. The I/O bandwidth at CPU can be formulated as:

$$I/O \text{ bandwidth at CPU} = (2 \times V_s) + (2 \times V_{ds}) + V_{I/O} \dots\dots\dots(4)$$

- Where  $V_s$ : Volume staged;
- $V_{ds}$ : Volume destaged;
- $V_{I/O}$ : Volume of I/O operations.

The data movement between the Data Handler (archive) and CPU gives the theoretical Processor <--> Data Handler throughput, which can be estimated by:

$$Theoretical \text{ Processor } <--> \text{ Data Handler throughput} = V_s + V_{ds} \dots\dots\dots(5)$$

**5.5.1.1.2.1 CERES**

The CERES daily average theoretical I/O bandwidth at CPU is shown in Table 5.5-4. Subsystems 4 and 5 are key contributors to both theoretical Processor <--> Data Handler throughput and I/O bandwidth at CPU. The CERES daily average I/O bandwidth for all processes is 7.18 MB/s. Similarly, the theoretical Processor <--> Data Handler throughput is 2.33 MB/s. Note that these numbers are based on raw AHWGP data for Standard Processing only.

**Table 5.5-4. CERES Daily Average Theoretical I/O Bandwidth at CPU and Processor <--> Data Handler Throughput**

Process	Volume staged (MB/day)	Volume destaged (MB/day)	I/O (MB/day)	Theoretical Processor <--> Data Handler throughput (MB/s)	I/O bandwidth at CPU (MB/s)
CERES 1aA	87	714	852	0.009	0.028
CERES 1aT	87	714	852	0.009	0.028
CERES 1bA	87	714	852	0.009	0.028
CERES 2aA	81	664	792	0.009	0.026
CERES 2aT	194	199	358	0.005	0.013
CERES 3aA	3	20	25	0.000	0.001
CERES 3aT	3	20	25	0.000	0.001
CERES 3bTA	7	20	28	0.000	0.001
CERES 4aF	5,109	6,076	12,524	0.129	0.404
CERES 4bA1F	65,125	6,076	74,722	0.824	2.513
CERES 4bA2F	65,125	6,076	74,722	0.824	2.513
CERES 5aF	3,819	5,481	10,565	0.108	0.338
CERES 5aV	616	884	1,704	0.017	0.054
CERES 5cAF	3,819	5,481	10,565	0.108	0.338
CERES 5cAV	616	884	1,704	0.017	0.054
CERES 6aA	5,481	99	6,820	0.065	0.208
CERES 6aT	5,481	99	6,820	0.065	0.208
CERES 7aA	424	263	697	0.008	0.024
CERES 7aT	424	263	697	0.008	0.024
CERES 7c	85	39	126	0.001	0.004
CERES 8aA	245	11	257	0.003	0.009
CERES 8aT	245	11	257	0.003	0.009
CERES 8c	245	11	257	0.003	0.009
CERES 9aAF	3,819	50	5,134	0.045	0.149
CERES 9aTF	3,819	50	5,134	0.045	0.149
CERES 10aA	311	17	328	0.004	0.011
CERES 10aT	311	17	328	0.004	0.011
CERES 10bTA	356	17	373	0.004	0.013
CERES 11a	9	0	18	0.000	0.000
CERES 12aF	32	252	334	0.003	0.010
Total	166,064	35,222	217,871	2.330	7.181

### 5.5.1.1.2.2 MISR

As shown in Table 5.5-5, the daily average I/O bandwidth at CPU for all MISR processes is 18.79 MB/s. The theoretical Processor <--> Data Handler throughput is 6.26 MB/s. Again, recall that these numbers are calculated based on raw AHWGP data for Standard Processing only.

**Table 5.5-5. MISR Daily Average Theoretical I/O Bandwidth at CPU**

Process	Volume staged (MB/day)	Volume destaged (MB/day)	I/O (MB/day)	Theoretical Processor <--> Data Handler throughput (MB/s)	I/O bandwidth at CPU (MB/s)
MISP1A	45,922	62,118	108,040	1.250	3.75
MISP1B1	55,332	56,971	112,303	1.300	3.90
MISP1B2I	110,548	63,293	173,841	2.012	6.04
MISP2ASI	71,891	6,989	78,866	0.913	2.74
MISP2TCI	64,163	4,002	68,150	0.789	2.37
Total	347,855	193,372	541,198	6.264	18.79

### 5.5.1.1.2.3 MOPITT

The MOPITT theoretical Processor <--> Data Handler throughput and I/O bandwidth at CPU shown in Table 5.5-6 are small compared to MISR and CERES.

**Table 5.5-6. MOPITT Daily Average Theoretical I/O Bandwidth at CPU and Processor <--> Data Handler Throughput**

Process	Volume staged (MB/day)	Volume destaged (MB/day)	I/O (MB/day)	Theoretical Processor <--> Data Handler throughput (MB/s)	I/O bandwidth at CPU (MB/s)
MOPL1	255	356	614	0.007	0.021
MOPL1Qi-D	356	10	366	0.004	0.013
MOPL2-E	344	185	589	0.006	0.019
MOPL2Qi-D	175	10	185	0.002	0.006
MOPL3	11	12	23	0.000	0.001
MOPL3Qi-F	11	1	12	0.000	0.000
Total	1,152	575	1,789	0.020	0.061

### **5.5.1.2 Dynamic Analysis Using ECS Systems Performance Model**

The ECS Systems Performance Model is used to dynamically simulate instrument processing at each of the DAACs. The AHWGP data (November 1994 baseline) served as input to the model. The model calculations should be viewed as preliminary. The AHWGP data are subject to change.

#### **5.5.1.2.1 Assumptions**

The following assumptions are made for the dynamic model simulations:

- Networks are not constrained;
- Disk storage is not constrained;
- No reprocessing;
- Data sets are not organized in the archive;
- No waiting storage (waiting storage may lower network needs);

#### **5.5.1.2.2 Dynamic Analysis for CERES**

Based on the number of instantiations (see Table 5.2-1) of processes, CERES processing is episodic. Processes are activated 24 times a day, some once a day, once a week and once a month. The ECS Systems Performance Model is used to simulate CERES processing to make a dynamic assessment of process completion times, processing resource usage and Processing <--> Data Handler throughput. A 300-MFLOP processor with a processor efficiency factor of 0.25 is used for the simulation. The model simulation is made for a 3-week period. During this 3-week period, there are days when there are daily, weekly and monthly processes activated on the same day.

##### **5.5.1.2.2.1 CERES Process Completion Times**

There are 30 CERES processes from various subsystems. Table 5.5-7 illustrates the model simulated process completion times (minimum, maximum, average and the standard deviation for each CERES process. On the average processes belonging to CERES Subsystems 4 and 5 take 300-3800 minutes to complete.

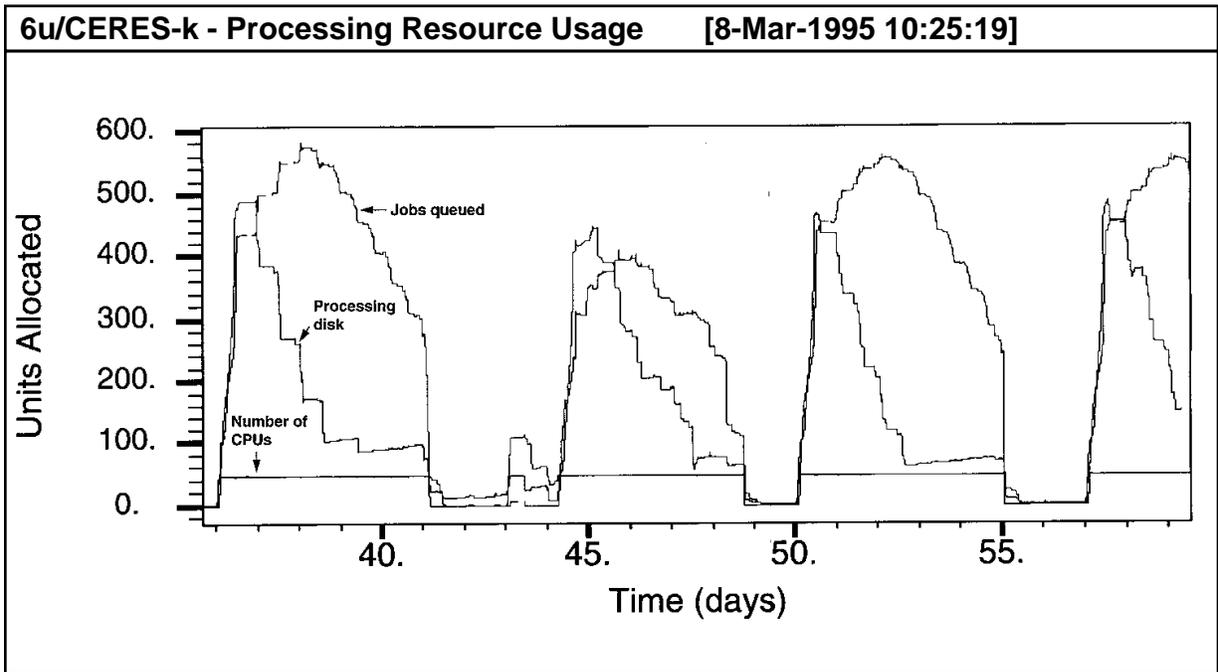
**Table 5.5-7. CERES Process Completion Times**

Processes	T <sub>max</sub> (minutes)	T <sub>avg</sub> (minutes)	T <sub>stddev</sub> (minutes)	T <sub>min</sub> (minutes)
CERES1aT	3999.70	1198.99	1489.89	4.62
CERES1aA	4001.18	1198.75	1490.87	4.62
CERES1bA	3998.31	1197.86	1490.48	4.62
CERES 2aT	1901.73	333.47	625.69	0.84
CERES 2aA	1900.88	332.12	624.44	0.84
CERES 3aT	30.97	20.73	10.23	10.50
CERES 3aA	30.92	20.71	10.21	10.50
CERES 3bTA	41.74	31.37	10.37	21.00
CERES 4aF	553.89	341.48	224.08	7.56
CERES 4bA1F	3686.07	1937.33	738.02	756.00
CERES 4bA2F	3686.07	1937.33	738.02	756.00
CERES 5aV	4736.46	2751.47	1490.78	593.88
CERES 5cAV	4682.16	3810.34	1033.62	593.88
CERES 5aF	4743.85	2698.90	1500.10	593.88
CERES 5cAF	4752.16	3830.15	1017.19	593.88
CERES 6aT	4124.75	2439.81	829.16	1.09
CERES 6aA	1887.48	658.09	604.37	1.09
CERES 7aT	151.56	151.56	0.00	151.56
CERES 7aA	151.56	151.56	0.00	151.56
CERES 7c	302.40	302.40	0.00	302.40
CERES 8aT	50.40	50.40	0.00	50.40
CERES 8aA	50.40	50.40	0.00	50.40
CERES 8c	2656.48	2656.48	0.00	2656.48
CERES 9aTF	4151.08	2118.31	1501.38	1.09
CERES 9aAF	4238.91	3248.34	1019.50	1.09
CERES 10aT	54.60	54.60	0.00	54.60
CERES 10aA	54.60	54.60	0.00	54.60
CERES 10bTA	2873.72	2873.72	0.00	2873.72
CERES 11a	4074.78	3739.46	531.43	2820.01
CERES 12aF	4004.68	1200.62	1486.13	8.40

**5.5.1.2.2.2 CERES Processing Resource Usage**

The episodic nature of CERES processing is clearly evident in the trace of Processing resource usage shown in Figure 5.5-2. For a maximum of 48 processors (as shown in Table 5.5-8), CERES processing resources (CPU and disk) show clear peaks when daily, and/or weekly and/or

monthly processes coincide. During these periods the number of jobs queued also increase, dropping off more slowly with time. Data product chains are responsible for the slow drop off of jobs queued. The short periods of lower activity is related to the maximum number of processors constrained in the model. Increasing the maximum number of processors will increase the period of lower activity. This is also called "processing slack". A slack is essential during real time operations to account for sudden and unexpected down times.



**Figure 5.5-2. CERES Processing Resource Usage**

**Table 5.5-8. CERES CPU and Staging Disk Capacity from ECS Systems Performance Model**

Number of CPUs (peak 300 MFLOPs)		Staging disk capacity (GB)	
Average	Peak	Average	Peak
31.3	48.0	120.6	492.8

**5.5.1.2.2.3 CERES Processor <--> Data Handler Throughput**

The Processor <--> Data Handler throughput estimates the amount of data movement between the Processor and Data Handler (archive). The model simulation gives a Processor <--> Data Handler throughput of 7.76 MB/s. This contrasts with 2.33 MB/s obtained theoretically (see Table 5.5-4).

### 5.5.1.2.3 MISR

MISR processing was dynamically simulated for a period of 6 days starting on the fifth day of the month. A 300-MFLOPs (peak) mid-range processor with a processor efficiency factor of 0.25 is assumed for the simulation. Recall from Table 5.2-3 that MISR processing is performed by orbit. Each process is activated approximately 14.56 times a day.

#### 5.5.1.2.3.1 MISR Process Completion Times

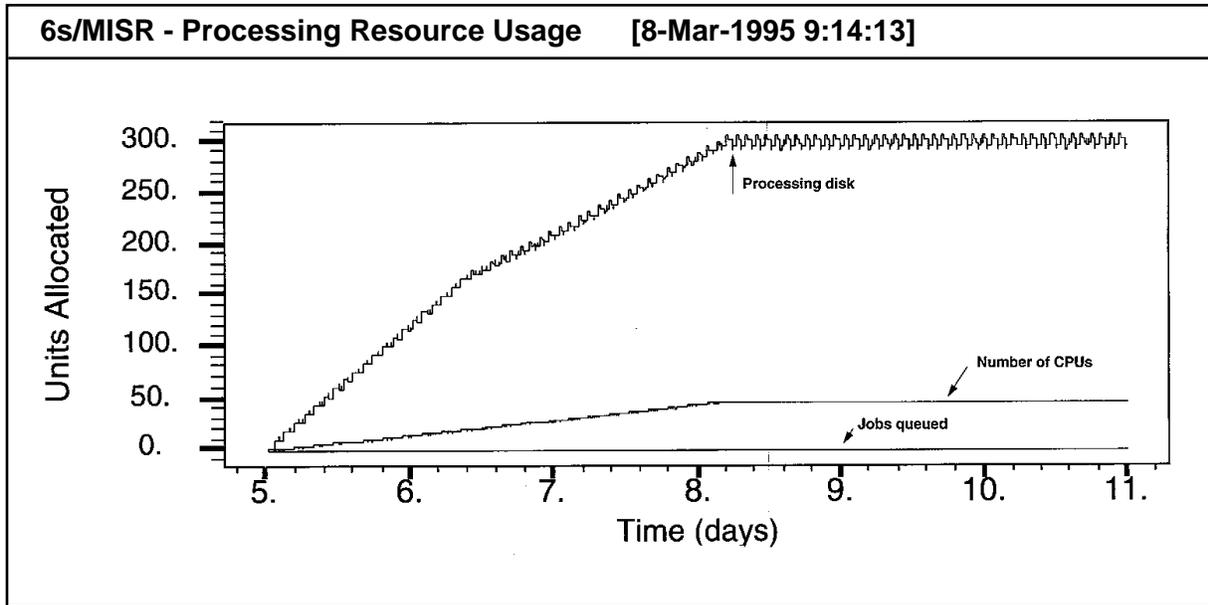
Since MISR processes are activated on an average of 14.56 times a day, the processing is not episodic unlike CERES. Table 5.5-9 shows the process completion times for each MISR process. The Level 1A and 1B1 processes (0 and 1) are estimated to run in less than 100 minutes. Process 2 (Level 1B2), however, can take on an average over 30 hours to complete. Processes 3 and 4 (Level 2 TC and AS) are estimated to take on an average 17 hours and 25 hours, respectively.

**Table 5.5-9. MISR Process Completion Times**

Processes	T <sub>max</sub> (minutes)	T <sub>avg</sub> (minutes)	T <sub>stddev</sub> (minutes)	T <sub>min</sub> (minutes)
MISP1A	52.66	52.66	0.00	52.66
MISP1B1	43.70	41.88	1.73	39.55
MISP1B2	1901.55	1901.55	0.00	1901.55
MISP2AS	1553.33	1553.33	0.00	1553.33
MISP2TC	1033.77	1033.77	0.00	1033.77

#### 5.5.1.2.3.2 MISR Processing Resource Usage

The dynamically simulated processing resource usage is illustrated in Figure 5.5-3 as processing disk capacity, number of processing CPUs and the number of jobs in the queue. Both input data staged for the L1 processes and the output produced by them exponentially fill up the processing disk. Higher level processes are activated only after all input data from lower level processes are available. Note that the number of CPUs required for processing gradually increases as higher level processes are activated. This simulation also represents a scenario where MISR processing is restarted after a total interruption. The simulation indicates that it can take up to three days for the production processing to attain steady state. Figure 5.5-3 and Table 5.5-10 indicate that an average of 46.3 processors (with 300 MFLOPs peak) are needed for MISR with an average staging disk volume of approximately 300 GB. With the processing load maintained by 46 processors, there are no jobs queued and waiting to be processed.



**Figure 5.5-3. MISR Processing Resource Usage**

**Table 5.5-10. MISR CPU and Staging Disk Capacity from ECS Systems Performance Model**

Number of CPUs (peak 300 MFLOPs)		Staging disk capacity (GB)	
Average	Peak	Average	Peak
46.3	47.0	299.7	305.6

**5.5.1.2.3.3 MISR Processor <--> Data Handler Throughput**

The model simulation gives a Processor <--> Data Handler throughput of 2.903 MB/s. This contrasts with 6.3 MB/s obtained theoretically (see Table 5.5-5). Since MISR processes are activated an average of 14.56 times a day and lower level products feed to higher level processes, the locality of particular data on the working storage is of importance. Therefore, the model has a scheduler that is intelligent enough to schedule processes whose data are already in the staging disk. Therefore, in the dynamic simulation, the number of bytes staged are much less than theoretical estimates.

**5.5.1.2.4 Dynamic Analysis for MOPITT**

MOPITT L1 and L2 processes are activated once a day. The L3 processes are activated once a week. A 100-MFLOPs (peak) workstation category processor is assumed for the simulation. The simulation is performed for a 10-day period. This time period is representative of daily and

weekly processing of all MOPITT products. The number of processing CPU is constrained to be 1 in the model run. A 100-MFLOP (peak) workstation-class processor with an efficiency factor of 0.25 is assumed for the simulation.

#### 5.5.1.2.4.1 MOPITT Process Completion Times

The process completion times of various MOPITT processes are illustrated in Figure 5.5-4. With the exception of MOPL2-E (process number 2) which is estimated to take 17 hours, MOPITT processes are estimated to take less than 1 hour to complete.

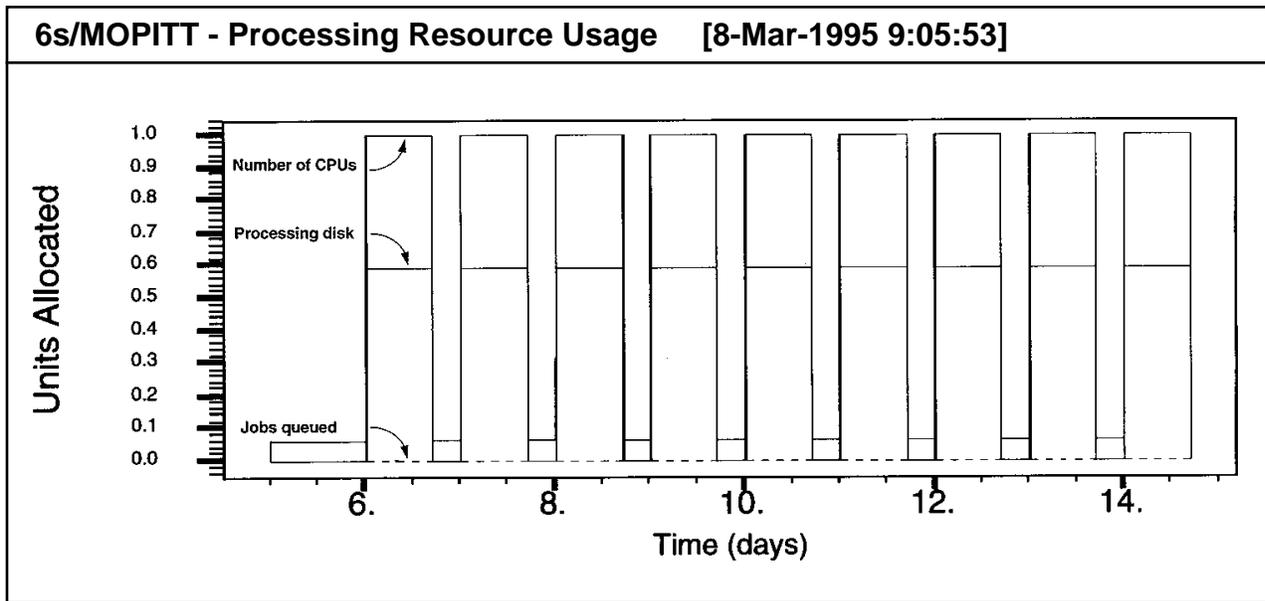


Figure 5.5-4. MOPITT Processing Resource Usage

Table 5.5-11. MOPITT Process Completion Times

Processes	T <sub>max</sub> (minutes)	T <sub>avg</sub> (minutes)	T <sub>stddev</sub> (minutes)	T <sub>min</sub> (minutes)
MOPL1	11.20	11.20	0.00	11.20
MOPL1Qi-D	1.07	1.07	0.00	1.07
MOPL2-E	1001.50	1001.50	0.00	1001.50
MOPL2Qi-D	1.19	1.19	0.00	1.19
MOPL3	18.86	18.86	0.00	18.86
MOPL3Qi-F	0.73	0.73	0.00	0.73

#### 5.5.1.2.4.2 MOPITT Processing Resource Usage

The L1-L3 processing is episodic and coincides with the data arrival rates. If all dependencies are met for each level of MOPITT processing, then it takes a little over half a day to process MOPITT data using 0.7 processors on an average. As shown in Table 5.5-12, the average daily processing disk space required is 0.4 GB with a peak requirement of 0.8 GB. The model also indicates that no jobs have been queued for processing. A process executes as soon as all the data dependencies have been satisfied. A residual disk space of 0.03 GB is required between process activations because there are permanent files that are required for processing different levels of MOPITT data.

**Table 5.5-12. MOPITT CPU and Staging Disk Capacity from ECS Systems Performance Model**

Number of CPUs (peak 100 MFLOPs)		Staging disk capacity (GB)	
Average	Peak	Average	Peak
0.7	1.0 (constraint limited)	0.4	0.8

#### 5.5.1.2.4.3 Processor <--> Data Handler Throughput

The model simulation gives a Processor <--> Data Handler throughput of 0.011 MB/s. This compares with 0.02 MB/s obtained theoretically (see Table 5.5-6). Because MOPITT processing is episodic and the staging and destaging volumes are small, the theoretical Processing <--> Data Handler throughput is a good estimate of the data flow between the Processing Subsystem and the Data Handler.

#### 5.5.1.3 Advantages

- Processing each instrument on a cluster will make administration easier
- Instrument requirements can be directly mapped to hardware;
- Product chains can be easily handled. If output from a Level 1 process is input to a Level 2 process, the files can remain in the staging area until the Level 2 process is ready to run. There can be substantial savings in the cost of moving data;
- Since the algorithms come from diverse instruments, it is believed that there may be special software and hardware requirements for each instrument. In this configuration, special hardware or software requirements can be localized on one cluster.

#### 5.5.1.4 Disadvantages

- An instrument's cluster can remain idle while other instruments' processing can have a backlog;

- Additional cluster infrastructure is necessary even if an instrument's processing requirements are small (e.g. MOPITT);
- Recovery time may be more in case of failure.

### **5.5.2 One Instrument's Products Per Cluster Except for Selected Products Requiring Major Processing Resources**

The requirements of each instrument (see Section 5.2) and resulting analyses (see Section 5.5) have yielded the following:

- It may be appropriate for MISR to be assigned to a high performance cluster. The number of activations per day combined with data volumes staged and destaged, and product chain dependencies may make it prohibitive to assign MISR processes to different clusters. A dynamic simulation is necessary for further understanding;
- CERES Subsystems 4 and 5 may be suitable for processing on a shared high performance cluster with MISR. Again, a dynamic simulation is necessary.

The dynamic analysis with the ECS Systems Performance Model will be performed during the CDR phase.

#### **5.5.2.1 Advantages**

- Special hardware/software requirements of multiple instruments can be localized.

#### **5.5.2.2 Disadvantages**

- For processing requiring major processing resources and having product chain dependencies, data may have to be moved to another cluster. This may unnecessarily increase data hops.

### **5.5.3 Multiple Instruments' Products Per Cluster**

This may apply to conditions whereby instruments with interdependent processing may be collocated. Both CERES and MISR require MODIS products generated at GSFC. However, the MODIS products required by CERES and MISR are different. Also, there is no interdependency among the three instruments. Therefore, the LaRC scenario is not ideal for analyzing this optimization alternative. During the CDR time period, a scenario at another DAAC will be analyzed with the ECS Systems Performance Model.

#### **5.5.3.1 Advantages**

- Instruments dependent on one another can be collocated, thereby, minimizing data hops.

### **5.5.3.2 Disadvantages**

- For processing requiring major processing resources and having product chain dependencies, data may have to be moved from one cluster to another. This may unnecessarily increase data hops;
- It may be more difficult to map individual processing requirements to hardware selection;
- Each instrument's growth in processing requirements can have global consequences.
- Instrument unique hardware/software requirements cannot be localized. Duplication may be necessary that can drive up costs;
- For processes with product chain dependencies, data may have to be moved from one location to another. This may unnecessarily increase data hops;
- This alternative may optimize resources but can introduce additional complexities for the Planning and Data Processing Subsystems.

### **5.5.4 Any Instrument's Products on Any Cluster That Can Support it; Selected by Current Processing Load**

This option is a mix-and-match situation. The processing load will determine the cluster where a particular instrument's data will be processed. This alternative allows the use of a large supercomputer to process many instruments at a DAAC site.

#### **5.5.4.1 Advantages**

- Idle time can be minimized because jobs are processed depending upon the current processing load on a cluster.

#### **5.4.4.2 Disadvantages**

- Mapping requirements to hardware is more difficult;
- Each instrument's growth in processing requirements can have global consequences;
- Instrument unique hardware/software requirements cannot be localized. Duplication may be necessary that can drive up costs;
- For processes with product chain dependencies, data may have to be moved from one location to another. This may unnecessarily increase data hops;
- This alternative may optimize resources but can introduce additional complexities for the Planning and Data Processing Subsystems.

This page intentionally left blank.

## 6. Conclusions

---

The conclusions from this study listed below should be considered preliminary:

- For Release B and beyond, it is important that cluster optimization alternatives identified in this study be considered before selection of Data Processing hardware classes. The alternatives can potentially optimize communications, staging storage and ease operations management and control. A more detailed study with the ECS Performance Model is necessary. This will be performed during the CDR phase.
- When one instrument is processed per cluster:
  - a) mapping instrument requirements to hardware is easier;
  - b) handling product chains does require data movement from one cluster to another;
  - c) instrument specific software/hardware requirements can be localized;
  - d) one instrument's cluster can remain idle while other instruments' processing can have a backlog;
  - e) additional cluster infrastructure is necessary even if an instrument's processing requirements are small (e.g. MOPITT).
- When one instrument's products are generated on a cluster dedicated to it except for selected products requiring major processing resources:
  - a) instrument specific software/hardware requirements can be localized;
  - b) product dependencies can increase data hops from one cluster to another.
- When multiple instruments' are processed on a cluster:
  - a) instruments dependent on one another can be collocated to decrease data hops;
  - b) instrument specific hardware/software requirements cannot be localized. Duplication may be necessary which can increase costs;
  - c) to handle product chain dependencies, data may have to be moved from one cluster to another which may unnecessarily increase data hops;
  - d) may introduce additional complexities for the Planning and Data Processing Subsystems.
- When any instrument is processed on any cluster that can support it (selected based on the processing load):
  - a) idle time can be minimized because jobs are processed depending upon the current processing load on the cluster;
  - b) each instrument's growth in processing requirements can have global consequences at a DAAC site.

This page intentionally left blank.

# Abbreviations and Acronyms

---

AHWGP	Ad Hoc Working Group on Production
AI&T	Algorithm Integration & Test
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
BONeS	Block Oriented Network Simulation
CDR	Critical Design Review
CERES	Clouds and Earth's Radiant Energy System
CPU	Central Processing Unit
DAAC	Distributed Active Archive Center
ECS	EOSDIS Core System
EDC	EROS Data Center
EOSDIS	Earth Observing System Data Information System
EROS	Earth Resources Observation System
GB	Gigabytes
GSFC	Goddard Space Flight Center
HWCI	Hardware Configuration Item
I&T	Integration & Test
I/O	Input/Output
L1	Level 1
L2	Level 2
L3	Level 3
LAN	Local Area Network
LaRC	Langley Research Center
LIS	Lightning Imaging Sensor
MB	Megabytes
MFLOPS	Millions of Floating Point Operations Per Second
MFPOs	Millions of Floating Point Operations
MISR	Multi-Angle Imaging Spectroradiometer
MODIS	Moderate-Resolution Imaging Spectroradiometer

MSFC	Marshall Space Flight Center
PDR	Preliminary Design Review
PGE	Product Generation Executive
QA	Quality Assurance
RMA	Reliability, Maintainability, Availability
SDPS	Science Data Processing Segment
SDR	System Design Review
SDS	System Design Specification