

---

# End To End Modeling

## Randy Miller

---

10 January 1996

# What Do We Mean By 'End To End Modeling'?



**Definition:** End To End Modeling is modeling that provides the capability to analyze system response to essentially all system inputs, across all subsystems.

**Purpose:** The purpose of End To End Modeling is to

- Analyze subsystem interactions
- Examine system sensitivities to load assumptions

**Results:** The results of this modeling include

- Utilization estimates for system resources
- Event traces for system processing threads

# Methodology



Our End To End Modeling methodology is comprised of these steps:

**Scenario Definition:** Scenario definition identifies the set of assumptions to be made -- principally about system loads -- for a modeling study.

**Thread Analysis:** Thread analysis identifies the specific stimuli and responses to be studied for a scenario.

**Model Runs:** Model runs are performed using the assumptions which define the scenario of interest. The models may be *instrumented* to capture information about the behavior for the threads of interest.

**Synthesis:** The results of all of the modeling tools (static, dynamic, queuing) and benchmarking efforts are synthesized to provide system level and thread level results.

# Scenarios



**The scenarios were selected to investigate conditions of interest:**

- **System response under nominal load**
- **System response under peak load**
- **System response during excursions from baseline loads**

**A scenario provides a complete description of the system -- a snapshot in time and space:**

- **Push Load (Nominal or Peak)**
- **Pull Load (Nominal, Peak, or Excursion)**
- **Distribution**
- **DAAC (LaRC, EDC, and GSFC) and Calendar Quarter (3Q99, in all cases)**
- **System Design (Hardware Specification)**

# Scenarios - Push Load



## Production Processing:

- The production push load during a scenario will be comprised of the inter-related PGEs described in the Technical Baseline (AHWGP).
- In “nominal” push load scenarios, the day chosen does not feature stressful periodic product generation (e.g., weekly or monthly summary products).
- In “peak” push load scenarios, the day chosen features the most stressful periodic product generation.

## Reprocessing:

- Reprocessing is modeled following the Head Of Chains paradigm described at IDR.
- The amount of reprocessing (1X or 2X) load will be appropriate for each instrument for the calendar quarter selected.

# Scenarios - Pull Load



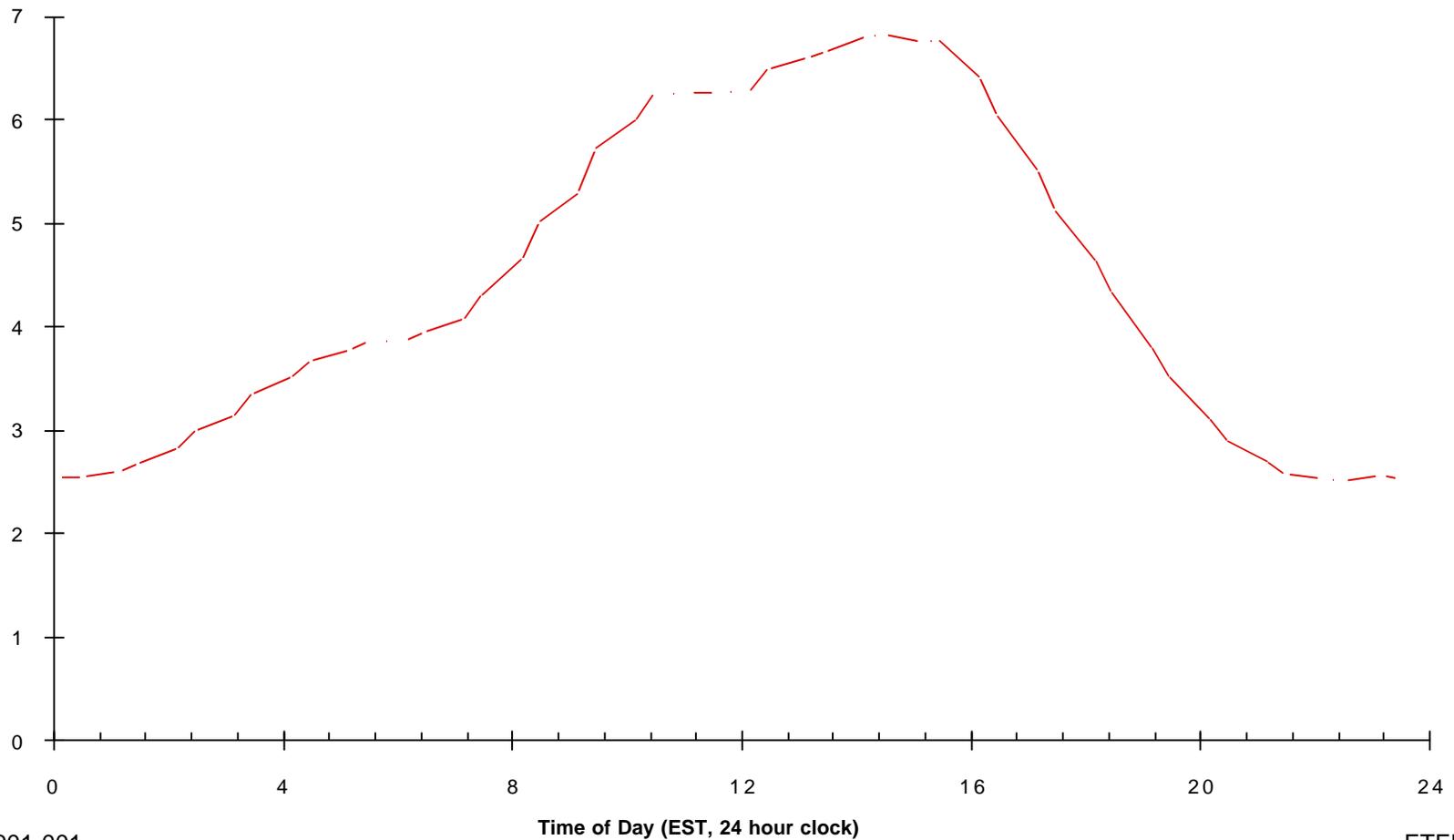
Two models of Pull Load are available:

- The User Model
- Functional and Performance Requirements Specification (F&PRS) Table 7-4

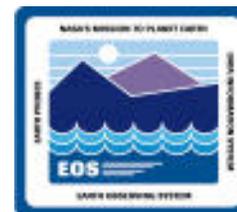
## User Model:

- The User Model is based on 27 scenarios developed by the Science Office.
- The User Model load follows a diurnal curve, with a peak to average ratio of approximately 1.5:1.
- The User Model predicts a distribution of the load among the DAACs.
- The User Model scenarios were initially mapped to 15 services, and more recently have been mapped to 64 services.
- The Dynamic Model currently uses the 15 service version of the User Model, which predicts approximately 3.7 archive transactions per minute. The 64 service version of the User Model predicts approximately 5.6 archive transactions per minute.

# User Model Load Versus Time Of Day



# Dynamic Model Basis Functions (2)



**Parameter:** DAAC to which request is directed.

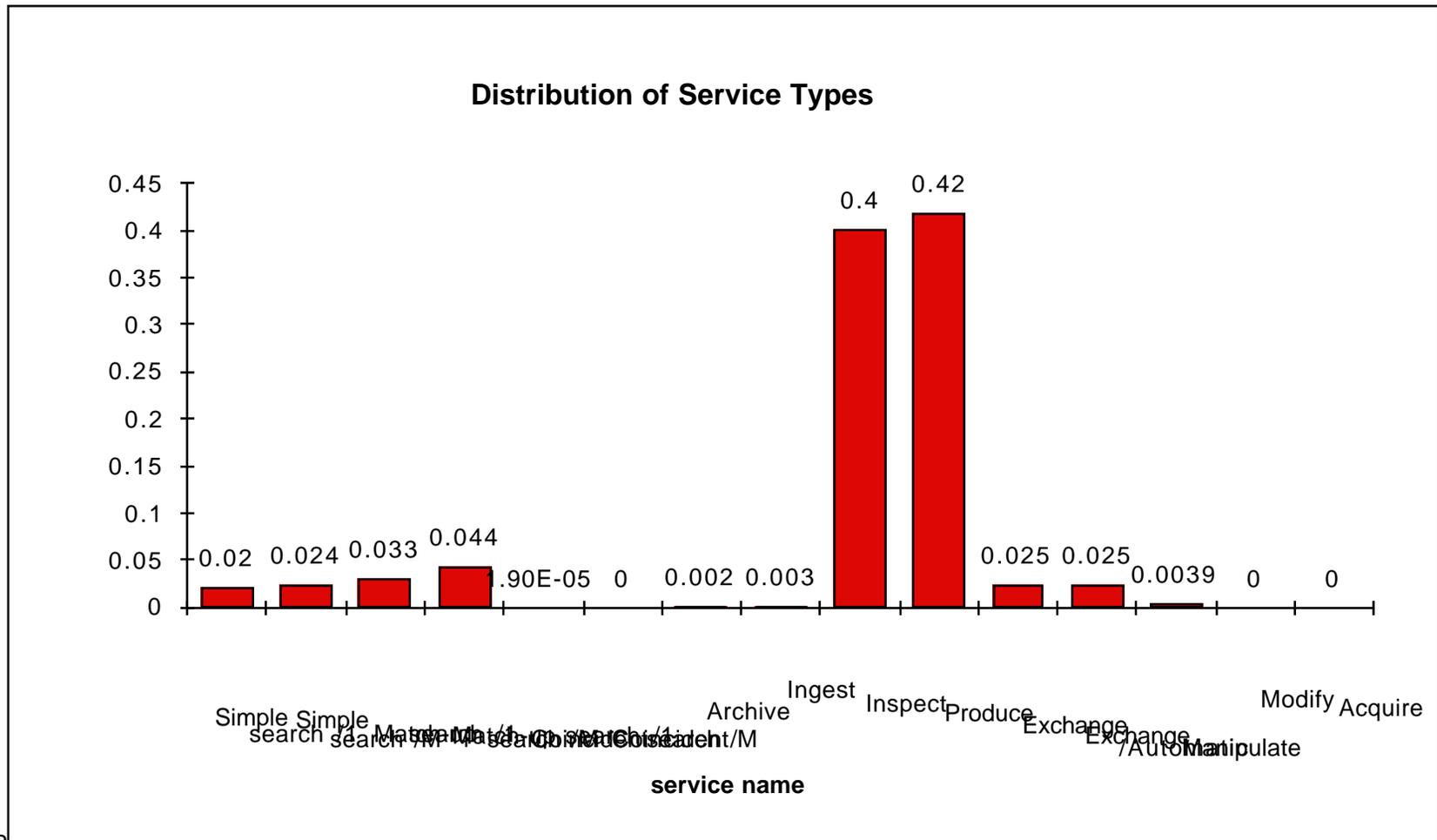
**Function:** Stochastic, probability per DAAC:

	<b>ASF</b>	<b>EDC</b>	<b>GSFC</b>	<b>JPL</b>	<b>LaRC</b>	<b>NSIDC</b>	<b>ORNL</b>
Probability	3.9%	48.4%	33.4%	1.7%	11.4%	0.6%	0.6%

**Rationale:** Best data available.

**Notes:** Probability developed from the User Model.

# User Model Distribution of Service Types



# Scenarios - Pull Load (Continued)



## F&PRS Table 7-4:

- Table 7-4 of the F&PRS specifies response time requirements and goals for 14 user operations, at specified operations per hour rates.
- The cumulative operations rate specified in Table 7-4 is 8.1 operations per minute, for a single DAAC.
- Approximately 10% of these operations access the archive to retrieve a single granule of data.

## Conclusions:

- The User Model and Table 7-4 of the F&PRS are difficult to rectify because they include different sets of operations.
- The Dynamic Model inputs will be updated to reflect the newer service definitions in the User Model, and the mapping of the User Model to the design.
- The operation per minute rates of the F&PRS will be used to drive the Queuing Model, using a peak to average ratio of 1.5:1 as suggested by the User Model.

# F&PRS Table 7-4



Session Category	Number of IMS Operations per Hour	Specific Operation
Log-on and Authorization	100	Account confirmation and authorization
Directory Search	80	Search by single keyword attribute
		Search by multiple keyword and time or space range check
Guide Search	40	Search for document by keyword
Inventory Search	120	Search one instrument by multiple keyword attribute with time or space range check (one DAAC)
		Search multiple instruments by multiple keyword attribute with time or space range check (one DAAC)
		Multiple DAAC inventory search by keyword attributes and time and/or space range check
Status Check (account or request)	60	Status of pending order or Data Acquisition Request
		Account status retrieval
Browse (for data selection)	50	Retrieve and begin to display standard pre-computed browse product
Document Search	10	Search 1000 document pages by keyword
Ordering Services	25	Local DAAC order submission and confirmation
		Remote DAAC order submission and confirmation
		Order cost estimate

# Scenarios - Distribution



## Distribution

- QA and Science User pull total 2X for the DAAC, unless an excursion from 2X in user pull is being investigated
- 2X defined by Technical Baseline

# Scenarios Selected For End To End Briefing



**Push**

**Pull**

	<b>Nominal</b>	<b>Peak</b>
<b>Nominal</b>	LaRC	
<b>Peak</b>		EDC
<b>Excursion</b>	GSFC EDC	GSFC EDC



# Thread Analysis

**Each thread describes a sequence of inputs to the system and the event traces as the system responds to the inputs.**

**The push and pull threads were selected to**

- **Cover as many subsystems as possible for each scenario;**
- **Represent inputs and responses of interest to as large an audience as possible.**

# Thread Analysis (Continued)



**Push Threads:** The push thread traces the receipt by the system of a data granule. The processing of the granule is tracked

- From Ingest to Data Server;
- From Data Server to Data Processing (for initiation of a PGE); and
- From Data Processing (via production of a Product) to Data Server.

**Ancillary transactions involving Planning, Management, and Communications will also be tracked.**

**The background push load during the execution of the push thread will include all of the PGEs appropriate for that scenario (DAAC, date, and time of day).**

# Thread Analysis (Continued)



**Pull Threads:** The pull thread traces a user session involving the following basic activities:

- Logging on;
- Querying the data dictionary;
- Querying for a list of documents;
- Retrieving a document for browse;
- Querying for a list of granules;
- Retrieving a browse product;
- Ordering of a set of granules; and
- Receiving the ordered granules.

# Thread Analysis (Continued)



Each activity of the pull thread will be expanded into the sequence of system actions required to perform that activity. For example, querying for a list of granules expands into the following interactions:

- Web Client to Web Server: Client sends the request.
- Web Server to Common Gateway Interface: Web server interprets the request and routes it to the common gateway, en route to the DIM.
- Common Gateway Interface to DIM: The common gateway interface routes the query to the DIM.
- DIM to LIM: The DIM routes the query to the LIM for the local portion of the query.
- LIM to DSS: The LIM routes the query to the Data Server.

. . . and the Data Server's results are routed back to the user.

# Thread Analysis (Continued)



**In the example, it was assumed that the client was web based, and the query required access to multiple DAACs. Assumptions will be made about the fraction of web clients versus Motif clients, the number of single-DAAC queries versus multiple-DAAC queries, and whether or not the queries access the VO Gateway.**

**The overall user load will be built up from the activities included in the pull thread, the frequencies for these activities identified in the F&PRS, and the assumptions cited above about client type, single/multi DAAC access, and VO Gateway access.**

# LaRC

## Nominal Push/Nominal Pull



### Motivation:

LaRC push processing is characterized by relatively small day to day variations in load, requiring relatively less reserve for peak push processing. On a typical day in this environment, what does system response and resource utilization look like?

### Thread Descriptions:

- Push: A CERES AM-1 Level 0 granule is received on a day when monthly CERES products are not being produced.
- Pull: The user logs on at 7:00 A.M. (nominal user loading).

# EDC

## Peak Push/Peak Pull



### Motivation:

Peak push processing at EDC, driven by the execution of periodic MODIS L3/L4 PGEs, places a significant stress on the DAAC's processing reserves. On a day with peak push requirements, how responsive is the system at EDC to user requests during the peak period of normal daily user activity?

### Threads:

- Push: A MODIS Level 2 granule (MOD04\_L2\_G) is received from GSFC that enables the scheduling and execution of a daily MODIS Level 3 PGE (MOD04:L3:DY:G). This occurs on a day when the MODIS MOD09:L3:16DY:G PGE is being executed.
- Pull: The user logs on at 3:00 P.M. (peak user loading).

# GSFC

## Nominal Push/Pull Excursion



### Motivation:

Normal push processing at GSFC has modest day to day variation due to the production of MODIS L3/L4 products; this creates a requirement for some reserve capacity to meet peak push processing requirements. On a day with nominal push requirements, what is the capability of this reserve capacity to satisfy excursions from normal pull demand?

### Threads:

- Push: A MODIS Level 0 granule (MOD\_L0\_G) is received on a day when periodic MODIS L3/L4 products are not being produced.
- Pull: The user logs on at 3:00 P.M. (peak user loading).

### Excursion:

Subscription distribution is varied from 1X to 4X while ad hoc distribution is held at 1X.

# EDC

## Nominal Push/Pull Excursion



### Motivation:

Normal push processing at EDC has significant day to day variation due to the production of MODIS L3/L4 products; this creates a requirement for large reserve capacity to meet peak push processing requirements. On a day with nominal push requirements, what is the capability of this reserve capacity to satisfy excursions from normal pull demand?

### Threads:

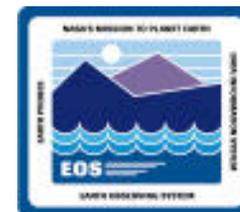
- Push: A MODIS Level 2 granule (MOD04\_L2\_G) is received from GSFC that enables the scheduling and execution of a daily MODIS Level 3 PGE (MOD04:L3:DY:G). This occurs on a day when periodic MODIS L3/L4 products are not being produced.
- Pull: The user logs on at 3:00 P.M. (peak user loading).

### Excursion:

- Ad hoc distribution is varied from 1X to 4X (by increasing the number of user operations per unit of time) while subscription distribution is held to 1X.

# GSFC

## Peak Push/Pull Excursion



### Motivation:

Peak push processing at GSFC, driven by the execution of periodic MODIS L3/L4 PGEs, places a moderate stress on the DAAC's processing reserves. On a day with peak push requirements, how does the system respond to an excursion from normal distribution demand?

### Threads:

- Push: A MODIS Level 0 granule (MOD\_L0\_G) is received. This occurs while data sets are being staged for the execution of the MODIS MOD06:L3:MN:G PGE. (MOD06:L3:MN:G is an I/O intensive PGE which stresses GSFC networks, staging disk, and disk I/O devices.)
- Pull: The user logs on at 3:00 P.M. (peak user loading).

### Excursion:

Subscription distribution is varied from 1X to 4X while ad hoc distribution is held at 1X.

# EDC

## Peak Push/Pull Excursion



### Motivation:

Peak push processing at EDC, driven by the execution of periodic MODIS L3/L4 PGEs, places a significant stress on the DAAC's processing reserves. On a day with peak push requirements, how does the system respond to an excursion from normal distribution demand?

### Threads:

- Push: A MODIS Level 2 granule (MOD04\_L2\_G) is received from GSFC that enables the scheduling and execution of a daily MODIS Level 3 PGE (MOD04:L3:DY:G). This occurs on a day when the MODIS MOD09:L3:16DY:G PGE is being executed.
- Pull: The user logs on at 3:00 P.M. (peak user loading).

### Excursion:

- Ad hoc distribution is varied from 1X to 4X (by increasing the number of user operations per unit of time) while subscription distribution is held to 1X.

# Model Runs



**New model runs will be performed, incorporating changes to the Technical Baseline (when available) and portraying the study scenarios:**

- **Static Model runs reflecting the new Technical Baseline (when available) will be performed to establish initial parameters and configurations for dynamic and queuing model runs.**
- **Dynamic Model runs will be performed to provide detailed Ingest, Data Server, Science Processing, and Distribution behavior.**
- **Queuing Model runs will be performed to provide Data Manager, Management, Communications, Planning, and Scheduling behavior.**

# Synthesis



**Results from all of the End To End Modeling activities -- Scenario Definition, Thread Definition, and Modeling -- will be combined to synthesize System Activity Level and Response Time results:**

- **System Activity Level results will be extracted from each model and collected together from each scenario into a single spreadsheet.**
- **Response Times for the events in the threads will be calculated from the model results. The threads will be mapped in a spreadsheet to event traces. Parameters for analyzing the system response time for each event will be extracted from the dynamic and queuing models. The spreadsheet will sum the timing components of the event trace to provide the total system response time for each step of the thread.**