A Performance Oriented Design Methodology for Large-Scale Distributed Data Intensive Information Systems

Daniel A. Menascé, Hassan Gomaa, and Larry Kerschberg

Abstract

The Earth Observing System (EOS) Data and Information System (EOSDIS) is perhaps one of the most important examples of large-scale, geographically distributed, and data intensive systems. Designing such systems in a way that guarantees that the resulting design will satisfy all functional and performance requirements is not a trivial task. This paper presents a performance-oriented methodology to design large-scale distributed data intensive information systems. The methodology is then applied to the design of EOSDIS Core System (ECS). Performance results, based on queuing network models of ECS are also presented.

Keywords

EOS, EOSDIS, ECS, performance modeling, complex systems, queuing networks, data intensive information systems.

I. Introduction

One of the most important examples of information systems that are large-scale, geographically distributed, and handle very large volumes of data is the Earth Observing System (EOS) Data and Information System (EOSDIS). EOS is a NASA program mission to study the planet earth. A series of satellites with scientific instruments aboard will be launched starting in 1997. They will collect data about the atmosphere, land, and ocean. An estimated terabyte of raw data will be sent to the earth every day. George Mason University (GMU) was one of three (the others were Berkeley and North Dakota) selected universities to develop an independent architecture for EOSDIS Core System (ECS). GMU put together an interdisciplinary team composed of Earth scientists, computer, and information scientists. The Earth scientists of our team came from GMU's Computational Science and Informatics Institute, from the University of Delaware, the University of New Hampshire, and from the Center for Ocean-Land-Atmosphere Studies (COLA) in Maryland. The authors of this paper were involved with the computer and information aspects of the architecture design. A methodology had to be developed to design such a complex system. This methodology is performance oriented to guarantee that the final design will satisfy the functional performance requirements of the system. The methodology is general and can be applied to the design of any large-scale distributed data intensive information system. After presenting the methodology, we discuss how it was applied to the design of ECS.

II. Large-Scale Distributed Data Intensive Information Systems

This section characterizes a large-scale distributed data intensive information system (LSS), provides the principles to be used when designing an LSS, and gives the logical architecture of an LSS building block (the LSS Node).

A. Characterization of an LSS

Figure 1 depicts the various components of a LSS. Such systems can be characterized as follows:

- *large number of users:* the number of potential users of an LSS can range from 100,000 to millions of users.
- *diverse user population:* users may include researchers studying a particular domain of science running complex simulation models (e.g., Earth scientists studying ocean circulation models), policy makers at governmental agencies (e.g., Department of Energy, Environmental Protection Agency), international organizations (e.g., The World Bank), private industries (e.g., the oil industry, timber industry), and K-12 students. Moreover, the users of an LSS are assumed to be spread through very large geographical areas.
- *diversity in user requirements:* as a consequence of the diversity in user population one may also expect to have a wide variation in user requirements. While some users may pose very simple queries to the system, other users may submit complex requests that may involve evaluating very complex scientific models or correlating several image files. User requests may also vary widely in terms of the amount of data requested (from a few bytes to hundreds of gigabytes). Different types of users may have different performance requirements and different categories of users may be assigned different priorities to ensure that their performance requirements are met.

D. A. Menascé (menasce@cne.gmu.edu) is with the Department of Computer Science, George Mason University, Fairfax, VA 22030-4444. Hassan Gomaa (hgomaa@isse.gmu.edu) and Larry Kerschberg (kersch@isse.gmu.edu) are both with the Department of Information and Software Systems Engineering at George Mason University. This work was partially supported by Hughes Applied Information Systems under Contract ECS-00010.

- *high data intensity:* raw data is expected to arrive at an LSS from one or more sources (e.g., instruments mounted on Earth orbiting satellites, particle accelerators) at very high rates (e.g., from terabytes to petabytes (10^{15} bytes) per day).
- *diversity in data types stored:* the holdings of an LSS are assumed to include a large variety of data types, such as raw data, large data sets resulting from applying complex algorithms to the raw data, images, metadata (i.e., data describing the data), free format text, and multimedia documents.
- *function distribution:* the different functions of an LSS should be implemented by components (LSS nodes) that are geographically distributed. These components are connected through one or more interconnected networks. Distribution is important to provide modularity, fault tolerance, and scalability to the design of an LSS.

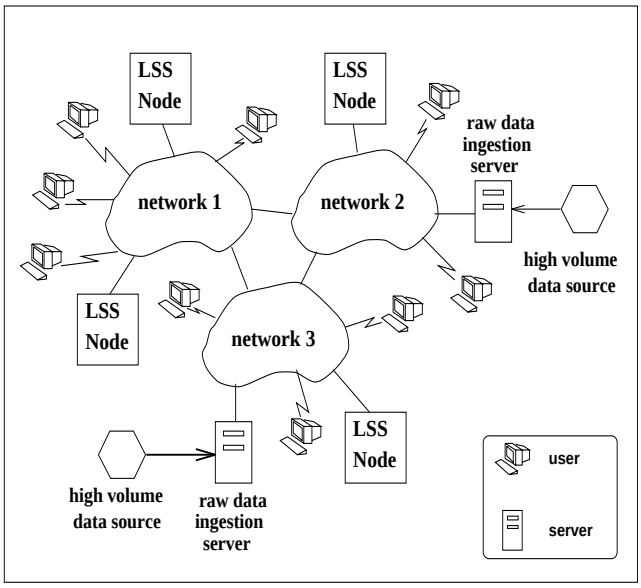

Fig. 1. A Large Scale Distributed Data Intensive Information System

B. Design Principles for an LSS

An LSS should adhere to the following principles:

- *P1. Location transparency*: users should be able to access any information object, including data products, metadata, and browse data, without having to know the physical location of these objects. This implies that data can migrate to achieve load balance, cope with failures, and improve performance without disrupting the users of the system.
- *P2. Modularity*: the architecture of an LSS should be composed of elements that can be configured to serve as nodes of different type, data processing, and storage capabilities, with the same underlying architecture.
- *P3. Minimization of User Connections*: the number of users that will access an LSS is very large. Thus, the system should allow for users to do as much local work as possible before a connection to the LSS is made.
- *P4. External processing capabilities:* networks of workstations (NOWs) are an attractive alternative to make available the idle cycles, unused main memory, and disk space of a collection of workstations, to parallel programs [1], [2]. An LSS should be capable of registering idle CPU cycles at user facilities and schedule computations using these cycles. There should also be a mechanism by which users of an LSS could pay for part of their usage through cycles.
- *P5. Separation of Functions:* the functions provided by an LSS should be divided into related groups and implemented by separate types of servers. This way, servers can be optimized to perform the functions they are best suited for.
- *P6. Scalability:* an LSS should be scalable to take into account different requirements that may exist at different stages of its lifecycle. Requirements may change as new data sources are incorporated into the system (e.g., new data collection satellites being launched) and as new users learn about the system. The system should be scalable on a selective basis. This means that if more I/O capacity at a given site is required, it should be possible to upgrade the I/O subsystem without necessarily impacting the processing and networking capabilities.
- *P7. Support for Heterogeneity:* an LSS should support many different types of processing paradigms to accommodate, as best as possible, the diversity of processing needs that may occur within a given complex application or in the collection of applications submitted to the system. Machines of different paradigms (e.g., SIMD, MIMD) as

well as heterogeneous architectures (e.g., Cray's T3D) should be allowed to coexist at the processing servers. The motivation for heterogeneity has been demonstrated in [20], [21], [22]. The Scheduler should take the heterogeneity into account when making its decisions. Significant work has been done in the area of metacomputing [6], [7] and on scheduling of parallel applications in heterogeneous environments [16], [18], [19].

- *P8. Minimization of Data Transmission:* Data transmission should be kept to the minimum possible level. This implies that data sets should be transmitted from source to destination with the minimum possible number of intermediate nodes. This principle is particularly important in a LSS where huge volumes of data are transported in and out of the system. A virtual client protocol was proposed [14] for minimizing data transfers in nested client/server interactions.

C. Logical Architecture of an LSS

The basic building block of an LSS is the LSS node shown in Fig. 2. The functions of an LSS are provided by a collection of interconnected LSS nodes.

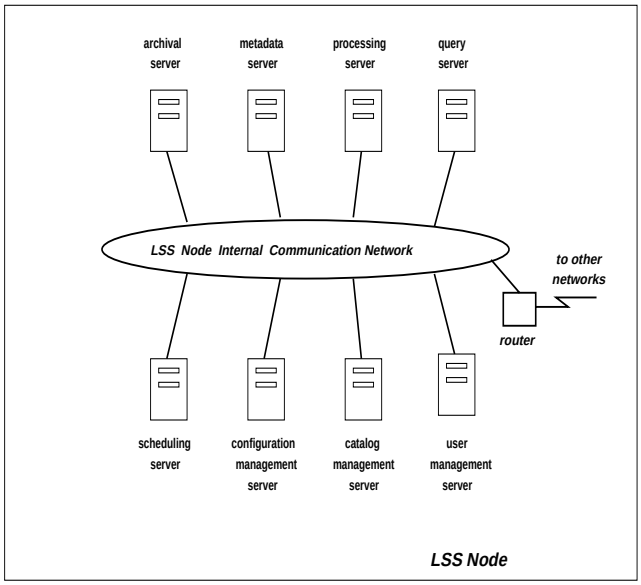

Fig. 2. Logical Architecture of an LSS Node

The basic logical architecture of such a node includes a collection of servers that implement the different functions of an LSS. Servers in an LSS node may act as clients with respect to other servers in the same or other LSS node. The main types of servers in an LSS node are:

1. archival server: handles the storage of all types of data in an LSS node. This type of server may be further specialized into archival servers of different types.
2. metadata server: manages the collection of metadata relative to the data managed by the archival server.
3. processing server: handles processing requests to transform data sets of one type into data sets of another type.
4. query server: manages the processing of both ad-hoc and pre-registered queries.
5. scheduling server: schedules the processing requests using both local and remote processing servers. The scheduling server may also use idle cycles located at user facilities. The set of scheduling servers in all LSS nodes, collectively implement a global scheduler.
6. configuration management server: monitors the operation conditions of the LSS node, collects statistics about the utilization of its various resources, and reconfigures the node when necessary to cope with failures and performance degradation.
7. catalog management server: maintains a directory of all objects managed by the LSS. The collection of all catalog managers collectively maintain a global directory of LSS objects. The catalog managers are used to locate LSS objects.
8. user management server: maintains information about registered users, their profiles, accounting and security information.

III. A Performance Oriented Design Methodology

Performance models play a crucial role in the design of complex information systems. They are useful to distinguish among a variety of alternatives, both good and bad, assess the impact of architectural choices, predict potential bottle-

necks, size hardware components, and evaluate if a proposed architecture will meet the performance requirements under the expected workload.

This section describes a performance oriented system design methodology. The main thrust of this methodology is to guarantee, by successive refinements, that the architecture meets performance goals set forth in the requirements analysis and specification phase. The methodology is better explained with the help of Figure 3.

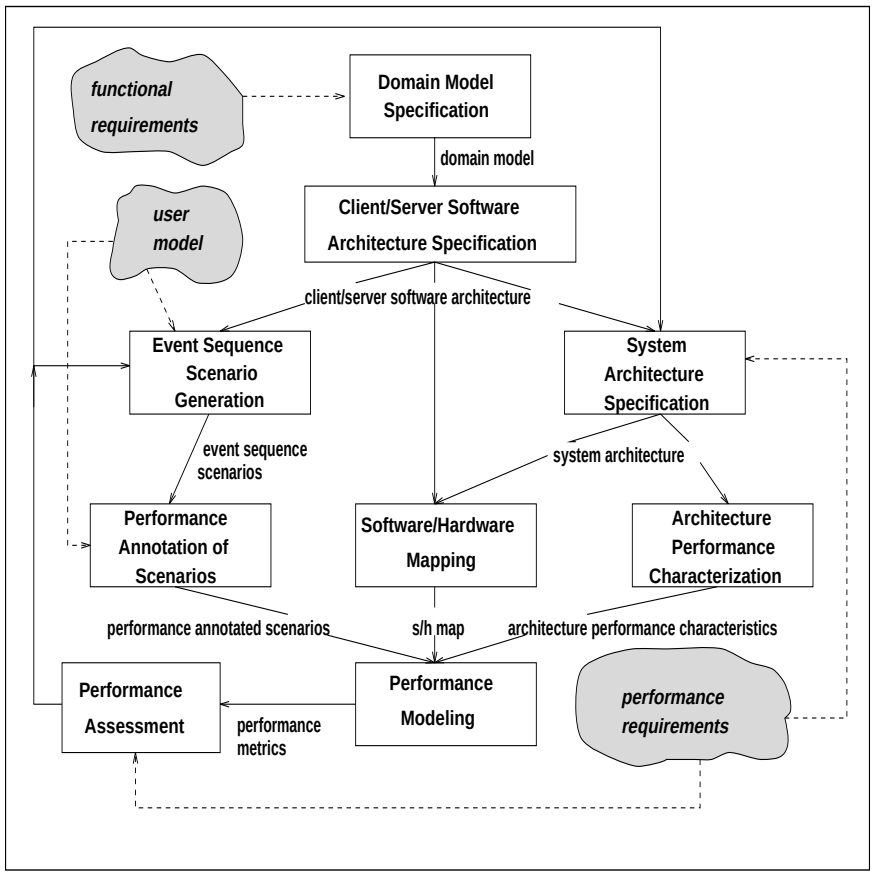

Fig. 3. Performance Oriented Design Methodology

There are three basic inputs to the methodology: functional requirements, performance requirements, and the user model. These three elements are shown as shaded clouds in Fig. 3. The functional requirements specify the functions to be performed by the system. The performance requirements specify the requirements on performance when executing any of these functions (e.g., maximum response time values, minimum throughputs). The user model describes the typical interactions between users and the system. The user model also provides quantitative information on the frequency with which users interact with the system, as well as the resource requirements per interaction (e.g., an Earth scientists studying ocean circulation models will typically browse twenty 3 MByte images and then will run an ocean circulation model that requires an average of 500 MFLOPs).

A domain model [9] is developed that reflects, at the functional level, the interaction of the main system components. The resulting domain model is then used to derive a client/server software architecture specification which depicts the message exchanges between clients and servers in the system. This client/server architecture is combined with the user model to generate event sequence scenarios showing each type of user interaction. These scenarios are further annotated with performance parameters such as request arrival rates, data volumes per request, server processing and I/O requirements per request. The client/server software architecture drives a first-cut at the system architecture. The client/server software architecture and the system architecture generate a software/hardware mapping that associates logical servers to physical elements such as processors and network segments. The components of the system architecture are assigned performance characteristics (e.g., network segment speeds, router latencies, I/O subsystem bandwidth, processor speeds). Then, the performance annotated scenarios, the software/hardware map, and the architecture performance characteristics are combined to generate input parameters to a performance model. The performance model is based on analytical methods to solve mixed (i.e., open/closed) queuing networks [15]. The outputs of the performance model include response times and throughputs for each type of request submitted to the system. An analysis of the results of the performance model reveals the possible bottlenecks. If the architecture does not meet the performance

objectives, architectural changes at the hardware and/or software level have to take place. These changes are guided by the outputs of the performance model. These changes are reflected back into the architecture and into the event sequence scenarios. Successive iterations guarantee that the final design meets the performance objectives. Since the design process is iterative, one starts with a first cut at the architecture and goes through successive refinements to meet the performance goals. These refinements may imply combining servers into a single physical computing element, changing the software architecture by creating additional servers, or changing the underlying hardware characteristics (e.g., internal network bandwidth, processing element speeds, and I/O subsystem rates).

The methodology just described was used by the authors in the design of an alternative architecture for EOSDIS Core System (ECS). The next section briefly describes EOSDIS and ECS. The remaining sections of the paper discuss how the methodology was applied to the specific design.

IV. EOSDIS and ECS

Raw data coming from the satellites is first received at the White Sands complex in West Virginia. After some initial level of calibration, it is sent for archival and further processing at a collection of eight centers called Distributed Active Archive Centers (DAACs). There are currently eight DAACs: Goddard Space Flight Center (GSFC), Langley Research Center (LaRC), EROS Data Center (EDC), University of Alaska at Fairbanks (UAF), University of Colorado (CU), Jet Propulsion Laboratory, (JPL), Marshall Space Flight Center (MSFC), Oak Ridge National Laboratory (ORNL), and the Consortium for International Earth Science Information Network (CIESIN). The raw data received by the DAACs is said to be Level 0 data. Level 0 data is used to generate Level 1 data, defined as reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary info. Environmental variables at the same resolution and location as the Level 1 data are derived to generate Level 2 data. A set of variables mapped on uniform space-time grid scales, with some consistency and completeness, are called Level 3 data. Finally, the model output or results from analyses of lower level data is known as Level 4.

About 500 NASA selected scientists will determine which data products are to be generated by ECS (standard data products). The facilities where these scientists are located are called SCFs (Science Computing Facilities). The remaining sections of this paper concentrate on EOSDIS Core System (ECS), and in particular on its Science Data Processing Segment (SDPS).

V. The Layers of the Architecture

The architecture of ECS can be conceptualized as being composed of fours layers as indicated in Fig. 4: the Application Layer, the ECS Services Layer, the Distributed Object Management Layer, and the Communications Management Layer. The figure shows how the functions and services of each layer are decomposed as well as how functions in one layer make use of services in the same or lower layer. The Applications and ECS Services layers are the only ones in which ECS concepts are known. The layers below are used to support the services provided by the two top layers. The functions of the four layers are explained in the following subsections.

A. The Application Layer

The Application Layer is composed of functions that are executed both at user facilities and at ECS sites and deal with *ECS objects*. ECS objects may be standard data products, metadata, browse products, algorithms, documents, or support data (e.g., calibration and engineering data). The application layer functions fall into the following categories:

- Browsing Functions: allow EOSDIS users to navigate through the web of ECS concepts to discover what ECS objects are available for retrieval or access. For example, an ECS user interested in studying the ozone layer may find out, by using the browsing functions, that EOSDIS contains information about atmospheric composition and that the MODIS instrument collects information about aerosol levels in the atmosphere.
- Query Functions: these are the functions used to formulate queries to retrieve ECS objects. For example, a user may use the query functions to request EOSDIS to retrieve all sea-surface temperatures where the strengths of the concurrent wind exceeds x Newtons/m^2. If the result of the query exceeds a certain pre-determined threshold in terms of volume of data to be transferred, the query functions must provide the user with an estimate of the volume (in bytes) of the result of the query. Also, the user should be presented with alternatives for delivering the result of the query: i) on-line delivery, ii) e-mail delivery, iii) e-mail notification, iv) magnetic tape, and v) CD-ROM. Costs as well as delivery time estimates are associated with each alternative to facilitate the users decision. Depending on the size of the result, some alternatives are not even considered. Users will also use the query functions to retrieve (order) standard data products. Through *subscription* functions, users may request EOSDIS to repeat a query at regular time intervals (e.g., every three hours). Users should be able to give conditions under which the results of a routinely repeated query are sent to him/her (ECS should allow for user provided "filters" that will be run against the result of a query to determine if the result should be sent to the user.) [*principle P8*]
- Data Access Functions: the data access functions constitute an API (Application Program Interface) through which programs running at user facilities can access ECS data sets at the data granule level. This allows users to run

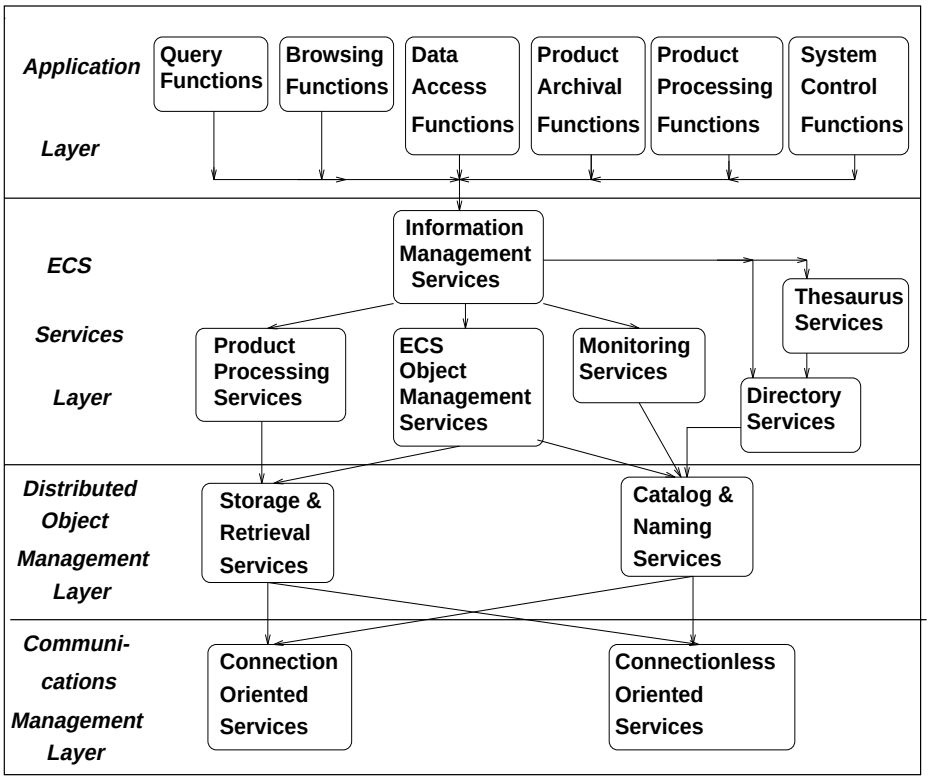

Fig. 4. ECS Multi-Layer Framework

scientific applications that require large volumes of data to be executed at their facilities without requiring that the entire collection of needed data sets be downloaded prior to the execution of the program. The API can also be used by components of ECS to request subsets of data products needed to process or reprocess standard data products.

- Product Processing Functions: these functions allow users to request the processing and reprocessing of standard products. Not every EOSDIS user should be allowed to request ECS to process or reprocess a particular standard product. Thus, these functions must ensure that the requesting user has the proper privileges to invoke these functions.
- Product Archiving Functions: through these functions, users with the right privileges can request ECS to archive a particular data product. These functions can be used to request the archival at determined sites of level 0 data or it can be used to request that higher level products generated within a DAAC be stored at a particular archival site.
- System Control Functions: most of the functions in this category are of interest to system administrators and can only be invoked by users with the right privileges. Examples of such functions include:
  - reporting on overall system performance and on individual component performance: these functions provide throughput and response time metrics as well as information on bottlenecks per workload type.
  - reporting on system and component failures and availability measures.
  - system reconfiguration functions to meet performance and availability requirements.
  - user registration: these functions allow for users to register with EOSDIS. ECS must keep user security (password and privileges) and profile information (used to optimize the processing of user requests) for each registered user.
  - billing functions: these functions generate bills to users for their use of EOSDIS.

B. ECS Services Layer

All services implemented by the ECS Services layer are independent of the location of the various servers that implement the services [*principle P1*]. This allows for EOSDIS resources (data and processors) to be relocated to comply with performance and availability requirements without the need to rebuild all the software running at the ECS Services layer. Two types of names (of ECS objects and ECS servers) are handled within the ECS Services layer: a *logical name* and a *mapped name*. The logical name for an object or server should be understandable by the users (e.g. "MOD27" as a name for the Ocean Productivity standard product as it is known by earth scientists). ECS objects at the Application Layer are only known by logical names. For each logical name there is a corresponding mapped name

which is used by the Distributed Object Management Layer to physically locate the object. The mapped name acts as a location independent object handle.

The services provided by the ECS Services Layer are:

- Information Management Services (IMS): these services provide the main interface through which the application layer provides query, browsing, data access, product archival, product processing, and system control functions to EOSDIS users. Depending on the type of user request, and based on directory information obtained from the directory services, other services within the ECS Services Layer may be invoked to perform the requested function. The IMS will typically request the result to be sent directly to the requesting user following a *Virtual Client Protocol (VCP)* aimed at reducing the amount of data transfer [*principle P8*] when nested client/server interactions exist. [14].
- ECS Object Management Services: these services support the implementation of the query, data access, and product archiving functions.
- Product Processing Services: these services support all the functions included in the set of Product Processing functions described in the Application layer.
- Directory Services: these services implement the mapping between logical names and mapped names. Every reference to an ECS object or ECS server within the ECS Services layer must use a mapped name. Thus, the Information Management Services need to use the Directory services to obtain the mapped name for a logical name.
- Monitoring Services: these are the services that support all system control functions defined in the Applications layer.
- Thesaurus Services: these functions are used to support the navigation through the web of ECS concepts provided by the Browsing functions at the Application layer. A more detailed discussion on thesaurus service can be found in [13].

C. Distributed Object Management Layer

At the interface between this level and the ECS services layer, objects are only known by their mapped names. The mapping between mapped names and location dependent names is implemented at this level. Note that only this mapping needs to be changed if ECS is reconfigured due to performance or component failure reasons. The services provided by the Distributed Object Management layer are:

- Naming and Catalog Services: these services implement the mapping between mapped names and location dependent names. Every time that a new ECS object is created at the ECS Services layer, a new mapped name is requested from the Distributed Object Management layer. This layer, through the Naming and Catalog services, allocates the object to an ECS physical location (ECS site). The ECS Services layer can provide guidance to the Distributed Object Management Layer as to the location of the newly created object. Since the ECS Services layer follows the principle of location transparency, it will tell the Distributed Object Management layer which other objects (given by their mapped names) should be co-located, if possible, with the new object.
- Storage and Retrieval Services: these services are used to store and/or retrieve an ECS object. The mapping between mapped and location dependent names is used to find the location of the object within EOSDIS.

D. Communication Services Layer

Two types of transport communication services are provided at this layer:

- Connection Oriented Transport Protocols: these services implement connection oriented protocols, such as TCP (Transmission Control Protocol). These are the types of services to be used for the transmission of large data streams when the sequence is to be preserved.
- Connection-less Transport Protocols: these services implement connection-less protocols, such as UDP (User Datagram Protocol) and they are used when short messages are to be exchanged by entities above the Communications Management Layer.

VI. Client Access to EOSDIS

The access mechanism to EOSDIS is explained with the help of Fig. 5. Users access EOSDIS through a software module called *EOSDIS Access Module (EOSX)* that must run at the user workstation. One of the main functions of EOSX is to connect to a *Client IMS (CIMS)* that may or may not reside at the user's desktop. If a user does not have a capable enough workstation to run a CIMS, he or she can always connect to a CIMS over a network.

The CIMS acts a server to EOSX but acts as a client to the *Server IMS* that runs at the DAACs. The Client IMS has the following components:

User Interface Manager (UIM): manages the interaction with the users connected to the CIMS.

User Profile Manager (UPM): manages the user profiles so that they can continue sessions from where they stopped previously, and keeps track of the preferred areas of interest for the user.

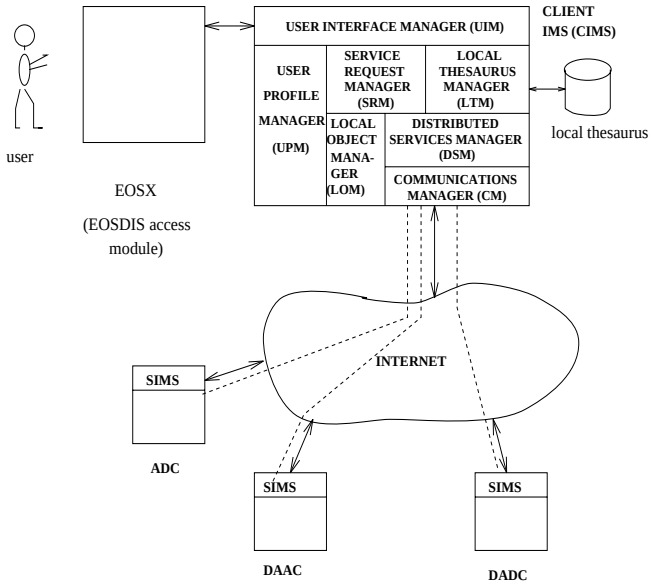

Fig. 5. Client Access to EOSDIS

Service Request Manager (SRM): handles requests to ECS services, to be passed on to the Distributed Services Manager, as well as requests to manipulate local objects (local objects may be browse products or data products previously retrieved from ECS and cached for the user at the CIMS).

Local Thesaurus Manager (LTM): maintains a local subset of the Global Thesaurus (see discussion below) maintained by the collection of all DAACs. The Local Thesaurus contains a description of a subset of the data available at EOSDIS along with the likely location of the data.

Local Object Manager (LOM): handles the management of all ECS objects (e.g., browse products, metadata, and data products) previously retrieved and cached for further use.

Distributed Services Manager (DSM): receives all requests for ECS services and generates appropriate requests to all ECS sites involved in the request. The DSM also acts as a transaction coordinator for user requests that involve multiple ECS sites. The requests generated by a DSM are sent to Server IMSs (SIMSs) at ECS sites.

Communications Manager (CM): establishes and manages all needed connections to ECS sites as requested by the Distributed Services Manager. If short requests are involved, connectionless protocols such as UDP should be used, while if large streams of data need to be exchanged, TCP connections should be used.

One of the main functions of the CIMS is to provide users with a subset of a Global Thesaurus (or Global ECS Web) implemented at the DAACs. The Global Thesaurus is a network of information nodes that describes ECS concepts available for users to search or navigate through. Stored with each information node is a list of ECS sites (DAACS, ADCs, or DADCs) that hold the actual data related to a particular information node. This Global Thesaurus has a set of information nodes, called entry points, that can be used to start a navigational process through the web.

A particular CIMS will typically contain a subset or sub-web of the Global Thesaurus, called the *Local Thesaurus*, that acts as a cache for the Global Thesaurus. This local cache stores the nodes and links most recently traversed by a user. Whenever a user attempts to traverse from an information node in his Local Thesaurus to an information node not in his Local Thesaurus, an *information node fault* is generated. As a consequence, a request is generated by the Local Thesaurus Manager (LTM) to any DAAC [1] to load a subgraph of the Global Thesaurus rooted at the information node that caused the fault. This request contains enough information on the users recent activity, profile, and storage limitations, to allow the Global Thesaurus Server at the DAAC to determine a proper subgraph that will potentially minimize the number of future information node faults.

Users should also be allowed to add their own local annotations to the Local Thesaurus. These local annotations are not reflected back to the Global Thesaurus.

Due to the principle of location transparency [*principle P1*], the location information contained in information nodes at the Local Thesaurus is not seen by the user but is used to determine where to send requests for ECS services. Note also that the Local Thesaurus is not updated every time that there is a change in the Global Thesaurus. This would be infeasible due to the huge number of users and therefore the large number of Local Thesauri in the system. Therefore, it is possible for an information node at the Local Thesaurus to be pointing to the wrong ECS site. This fact will be detected when a request to the ECS site is generated. Only then, the Local Thesaurus will be adjusted to reflect the

[1] It is assumed that the Global Thesaurus is replicated at all DAACs.

change. This procedure acts like an "optimistic cache coherence protocol" in the sense that it does not try to maintain the coherence of all Local Thesaurus caches, but rather operates under the optimistic assumption that the information is valid. It should be obvious that "cache invalidation" type of protocols would be prohibitive in the ECS environment.

Note that due to different types of caching mechanisms implemented at the CIMS (Local Thesaurus and Local Object Manager), connections and data transfer to and from ECS sites are kept to the minimum possible [*principle P3*].

VII. ECS Node Logical Architecture

The building block of the ECS architecture is called an *ECS node*, which is a collection of servers of different types. This is ECS's instantiation of an LSS node. ECS nodes can be configured to serve as DAACs, Auxiliary Data Centers, or even SCFs [*principle P2*]. ECS nodes are connected to the user community by a User Network (e.g. Internet/NII) and are connected to other ECS nodes by the ESN network.

An ECS node (see Fig. 6) has three main subsystems: ECS Object Management Subsystem (EOMS), Product Processing Subsystem (PPS), and Information Management Subsystem (IMS). Each of the main subsystems is implemented by a collection of servers to be described below. All servers can communicate with one another through an ECS Node local high speed network. Through this network, and ECS node is connected to the Internet and to the Earth Science Network (ESN). Note that Fig. 6 represents a logical not a physical design. In order to map the logical design into a physical design one would need to specify the mapping of the various servers into actual machines, the capabilities of these machines in terms of processing and I/O characteristics, and the topology and bandwidth characteristics of the ECS Node Local high speed network. More details on the specific functions of each server are given in [14].

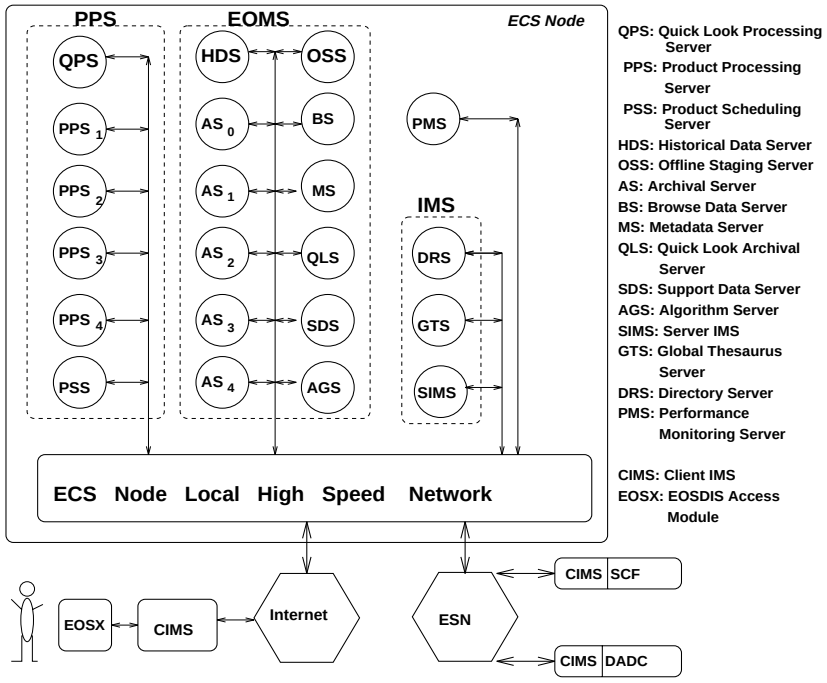

Fig. 6. ECS Node Architecture.

An ECS node may have any number of servers of each type. The actual number of servers is determined by the workload imposed on a given ECS node. A server may act as client for a server on the same or on a remote ECS Node. For example a Product Processing Server, acting as a client, may request a data product from several Archival Servers. Some of them could be local to the Product Processing Server and others could be remote. It is important to realize that servers are just logical concepts. The mapping between servers and actual machines determines the actual physical architecture of the system.

Although it is conceivable for all servers of an ECS node to be mapped into a single computer, this is not desirable for various reasons: need for redundancy, adequate performance, and the need to accommodate significantly different processing, communications, and I/O requirements.

VIII. A Performance Model for the ECS Architecture

Performance models of computer systems are used to predict how performance metrics such as throughput, response times, queue sizes, and component utilizations, vary as a function of the workload and system parameters. The analysis

discussed in this section is based on queuing network based analytic models [15]. We present here a brief overview of the concepts behind queuing network (QN) models as well as the terminology to be used.

A *queuing network (QN)* is a network of queues through which customers flow. Customers may have different meanings in different contexts. In the EOSDIS context, a customer may be a request to retrieve a browse image from a certain DAAC, or a request to generate a given standard product. A *queue* is composed of a *service center* and a waiting line of customers waiting to use the service center. Service centers may be used to represent a processing element, components of an I/O subsystem, or a communications network. Since customers may vary significantly in terms of the demands placed on the different service centers, one should aggregate all similar customers into groups called *customer classes*. All customers of the same class are represented by a set of values that represent the average demand on each service center over all customers of the class. In the context of EOSDIS, each different scenario generated from the user model may give rise to a different customer class. Some customer classes may be considered to be *open* in the sense that the customers arrive from outside the system, get served by a subset of the service centers, and leave. In this case, there is not limit on the number of customers in the queuing network. An example of an open class could be "browse image queries from K-12 students". Other classes may be considered to be *closed* in the sense that there is always a constant number of customers of this class in the queuing network. This type of class is used to represent work that is performed on a routinely and continuous basis. An example of this would be "processing of standard product MOD28".

Each customer class is specified by the following parameters:

- type: open or closed.
- intensity parameters: arrival rate of customers in the case of open classes and number of customers in the case of closed classes.
- service demands: total time spent per customer receiving service from each service center. Note that the time spent waiting to get access to the service center is not part of the service demand.

The total time spent by a customer in the queuing network is the response time for the customers' class. The response time is composed of two basic components: queuing time and service time. The queuing time, computed by queuing network based analytic models, is a function of the contention for access to the several service centers. The service time is a function of the total service demand placed on all resources. Another performance measure of interest is the throughput per customer class. This measures the average number of customer request completions per unit time. For open classes, the throughput is simply equal to the arrival rate. For closed classes, the throughput is computed by a queuing network based analytic model.

For each customer class, one must specify the required *service levels*, i.e., the upper and lower bounds on performance. Examples of service levels in the context of EOSDIS are:

- level 1 data must be made available within 48 hours of observation.
- levels 2 and 3 standard products should be made available within 96 hours of observation.
- A DAAC should be capable of generating quick-look products within 1 hour of receipt of necessary input for 1% of EOS instrument data.

Each user category may exhibit several important patterns of interaction with ECS. Each scenario is assigned to a class of customers in the queuing network model. Each scenario is mapped to the ECS architecture so that service demands can be obtained for the different architecture components and subsystems. Given that the architecture under consideration is a client-server based architecture, multiple time-line diagrams such as the one shown in Fig. 7 were used to map each scenario to the architecture. In this figure, an external request submitted to the Client IMS implies in three requests sent to Server IMS, one request sent to the Level 2 Metadata Server, and one request sent to the Level 2 Browse Data Server. Note that servers may act as clients when requesting service from other servers. The dashed arrows from a server (acting as a client) into a server indicate a request. These arrows are annotated with performance related data such as the probability that service is requested from this server, number of bytes involved in the request, and resource demand parameters related to the service requested from the server. The dashed arrows from a server to another (acting as a client) indicate replies from previous requests. These arrows are also annotated with the number of bytes sent back to the client.

An analysis of each scenario determines the average number of requests per server as well as the average service demand per server. This analysis also determines the average load imposed on the various communication subsystems of ECS.

A. Performance Model Parameters

The service demand parameters per server for each scenario (customer class in the queuing network model) are obtained as follows. Let

- $b(r)$: data volume to be retrieved by request r (in MBytes),
- $c(r)$: computational demand associated with request r (in millions of floating point operations),
- $\mathcal{C}_i$: computing speed of server i (in MFLOPS),
- IO_i: I/O bandwidth of server i (in MBytes/sec),

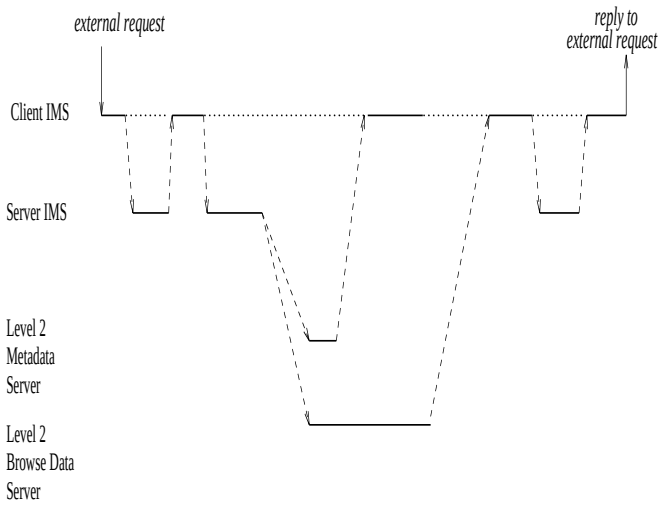

Fig. 7. Multiple Time-line Diagram

- $p_{i,j}(r)$: probability that request r is addressed to server j,
- $\mathcal{R}_i$: set of requests generated by client i,
- D_j^{proc}: processing service demand at server j,
- D_j^{IO}: I/O service demand at server j.

Then we can write that,

$$D_j^{\text{proc}} = \sum_{\forall i} \sum_{r \in \mathcal{R}_i} \frac{p_{i,j}(r) \times c(r)}{\mathcal{C}_i}$$

$$D_j^{\text{IO}} = \sum_{\forall i} \sum_{r \in \mathcal{R}_i} \frac{p_{i,j}(r) \times b(r)}{IO_i}$$

B. The Queuing Network Model

The performance model used for the ECS architecture is a mixed queuing network: some classes are open and some are closed. The data product generation classes are closed classes while queries (e.g., browse and metadata) are represented as open classes. Any server (in the client/server architecture) is represented by two devices in the QN: a computing device and an I/O device. The various communication networks are represented by load dependent devices in the QN. Figure 8 contains a diagram of the queuing network model used to represent the ECS architecture.

C. Performance Results for ECS Architecture

Several numerical examples are used in this section to evaluate the architecture of ECS. A baseline model using the values derived from the user model generated by the team of Earth scientists was evaluated first. Modifications analysis were then carried out to gauge the sensitivity of the architecture to different changes in the workload and architecture components. It should be noted that the results shown below do not represent actual ECS performance since, at the time of this study, we did not have enough data available. Some assumptions had to be made to compensate for missing values. At any rate, the results indicate the type of analysis that can be made with the performance models described here.

The following numerical data are considered in the baseline model:

- Local Area Network Bandwidth: 100 Mbps based on a 2-channel FDDI backbone.
- ESN bandwidth: 45 Mbps assuming T3 lines between the major DAACs.
- Processing capacity: 60 GFLOPS, based on the achieved performance for a CM-5 running the LAPACK benchmark [3].
- I/O bandwidth: 70 MBytes/sec.

For the purposes of this study we considered the following scenarios:

- *El-Nino and Southern Oscillations (ENSO)* [10].
- *World Ocean Circulation Experiment and the Tropical Ocean Global Atmosphere (WOCE/TOGA)* [11].
- *Global Ocean Observing System (GOOS)* [11].
- *Terrestrial Scenario (Land_use)* [12].

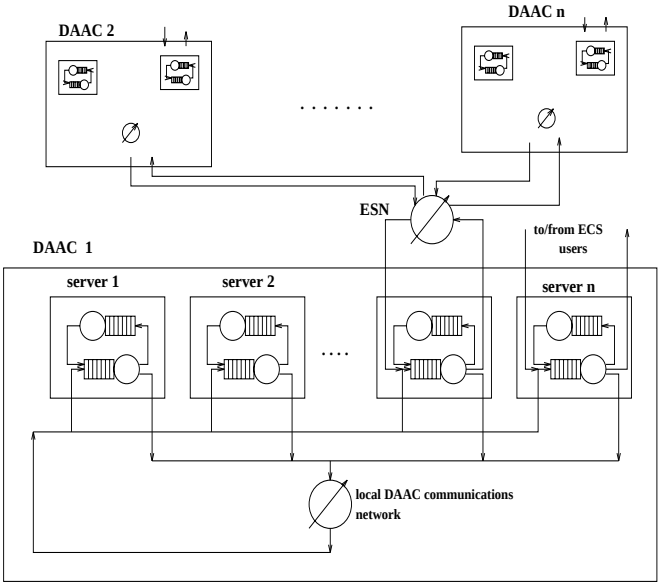

Fig. 8. Queuing Network Model of the ECS Architecture

- *Push scenario for Level 0 products (L0).* This scenario represents the workload imposed on ECS by the continued arrival, processing, and storage of level 0 products. The numerical data for this workload is derived from tables available in HAIS documents.
- *Push scenario for Level 1 and higher data products (L1-L4).* This scenario represents the workload imposed on ECS by the processing and storage of levels 1 through 4 data products. The numerical data for this workload is derived from tables available in HAIS documents.

We show here a few examples of the results obtained in our analysis. An extensive set of curves and tables can be found in [17]. The value of the workload intensity for the user scenarios which are not varying were chosen so that they represent a light load. For the open scenarios (ENSO, WOCE/TOGA, GOOS, and Land_use) the arrival rate was fixed at 0.1 requests/sec except when this is the varying parameter. For the L0 and L1-L4 scenarios the number of jobs in the system was fixed at 10 except when this is the varying parameter.

Figure 9 displays the impact of varying the arrival rate of ENSO requests on the response time of GOOS, Land_use, and L1-L4 classes. As it can be seen, response times smaller than 4 sec for Land_Use and GOOS are supported for arrival

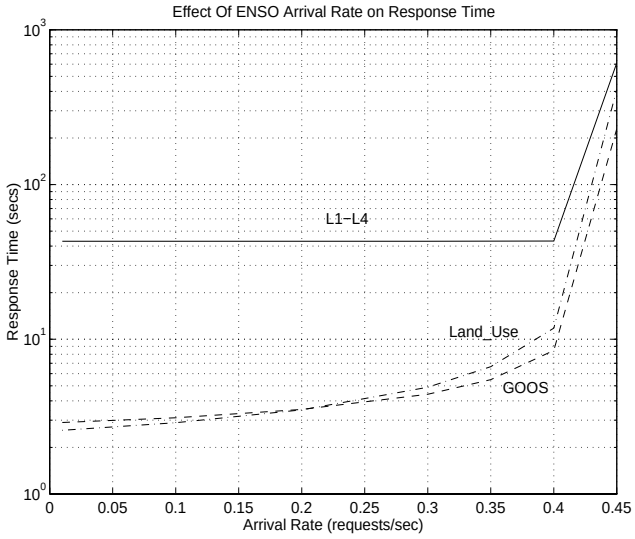

Fig. 9. Impact of ENSO Arrival Rate on GOOS, L1-L4, and Land_Use Scenarios.

rates of ENSO requests not exceeding 0.25 req/sec. After 0.4 req/sec for ENSO requests, the system saturates and the response times increases at a very fast rate. The L1-L4 workload is rather insensitive to ENSO requests until the onset

of saturation. The response time curve of the L1-L4 workload as a function of the number of jobs (see Fig. 10) grows to a certain saturation point since the throughput for closed classes is bounded by the maximum service demand [15]. Figure 11 investigates the impact of consolidating some of the DAACs. In particular, the loads of DAACs MSFC, JPL,

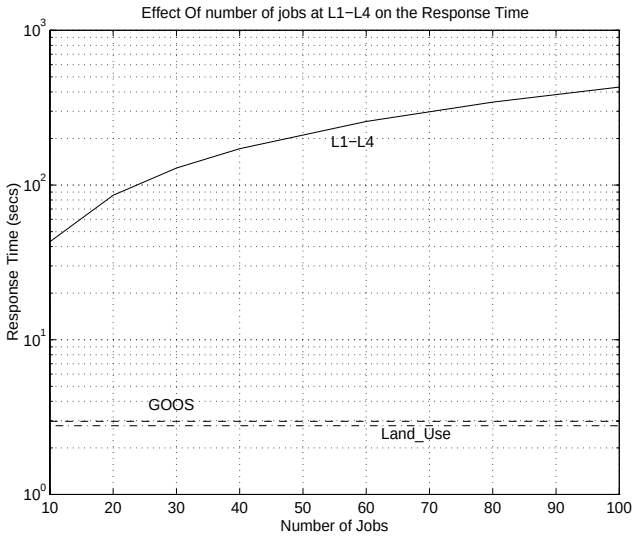

Fig. 10. Impact of L1-L4 workload intensity on GOOS, L1-L4, and Land_Use Scenarios.

UAF, and CU are assumed to be assigned to LaRC. This reduces the number of DAACs from 8 to just 5. The vertical axis in the figure is the response time ratio S defined as

$$S = \frac{\text{Response Time Under the Consolidated Scenario}}{\text{Response Time Under the Original Scenario}}$$

A value of S greater than 1 indicates that consolidating DAACs increases the response time. It should be noted that when DAACs are consolidated, there are two effects on performance: some servers will become more heavily utilized as a result of the additional load. This makes S to be greater than 1. On the other hand, for some scenarios, the network demand is reduced by the DAAC consolidation. This reduces the response time under the consolidated scenario and makes $S < 1$. As seen in Fig. 11, performance for the ENSO workload becomes twice as bad when the arrival rate of GOOS requests approaches 0.55 req/sec. The L0 workload is only slightly sensitive to DAAC consolidation.

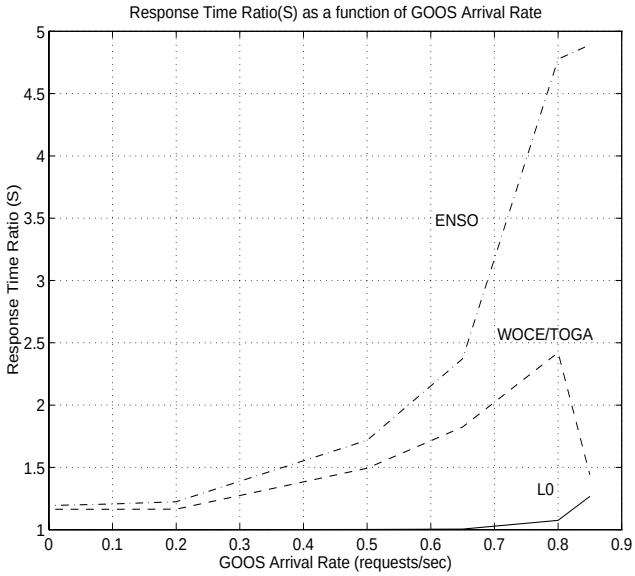

Fig. 11. Impact of GOOS Arrival Rate on ENSO, L0, and WOCE/TOGA Scenarios.

IX. Summary Remarks

A characterization of large-scale, distributed, and data intensive information systems was given. A general methodology to design such systems was discussed. The methodology is performance-oriented. Through the iterative use of performance models to carry out performance prediction, the system design is refined through successive steps. This guarantees that the final design will satisfy the functional and performance requirements. EOSDIS Core System was used a case study for the methodology.

Acknowledgements

The authors would like to acknowledge the many useful discussions they had with the independent architecture study group for EOSDIS ECS at GMU led by Menas Kafatos. In particular, the authors would like to thank Jim Churgin, Ferris Webster, Berrien Moore III, and Jim Kinter, for explaining them the different aspects of Earth science and user scientist requirements for EOSDIS. They would also like to thank Sudhanshu Killedar for writing the programs that implement the analytic model.

References

[1] Arpaci, R. H., A. Dusseau, A. M. Vahdat, L. T. Liu, T. E. Anderson, and D. A. Patterson, *The Interaction of Parallel and Sequential Workloads on a Network of Workstations*, Proceedings of the 1995 ACM Sigmetrics Conference, Ottawa, Canada, May 15-20, 1995.

[2] Cap, C. H. and V. Strumpen, *Efficient Parallel Computing in Distributed Workstation Environments*, Parallel Computing, 19:1221-1234, 1993.

[3] Dongarra, J., H-W. Meuer, and E. Strohmaier, TOP 500 Report 1993.

[4] Dongarra, Jack J., G. A. Geist, R. Manchek, and V. S. Sunderam, *Integrated PVM Framework Supports Heterogeneous Network Computing*, Oak Ridge National Lab Technical Report, January 1993.

[5] NASA, EOS Reference Handbook, Washington, DC, August 1993.

[6] Freund, R. F. Optimal selection theory for superconcurrency. *Proc. Supercomputing'89*, IEEE Computer Society/ACM Sigarch, Reno, NV. 1989, pp. 699-703.

[7] Freund, R. R., and Conwell, D. S. Superconcurrency, a form of distributed heterogeneous supercomputing. *Supercomputing Review*, June 1990.

[8] Gomaa, H., L. Kerschberg, and D. A. Menascé, *A Software Architectural Design Method for Large-Scale Distributed Data Intensive Information Systems* , submitted for publication, 1995.

[9] Gomaa, H., L. Kerschberg, and V. Sugumaran, *A Knowledge-Based Approach for Generating Target System Specifications from a Domain Model* , Proc. NASA Goddard Conference on Space Applications of Artificial Intelligence, May 1992. Also in Proc. IFIP World Computer Congress, Madrid, Spain, September 1992.

[10] Kinter, J., B. Doty, and J. Shukla, *Meteorological Scenarios*, The GMU ECS Federated Client-Server Architecture, Final Report, Chapter 5, Part II, George Mason University, August 31, 1994.

[11] Churgin, J. and F. Webster, *Oceanographic Scenarios*, The GMU ECS Federated Client-Server Architecture, Final Report, Chapter 4, Part II, George Mason University, August 31, 1994.

[12] Moore III, Berrien, *Terrestrial Scenarios*, The GMU ECS Federated Client-Server Architecture, Final Report, Chapter 3, Part II, George Mason University, August 31, 1994.

[13] Kerschberg, L., H. Gomaa, D. A. Menascé, and J. P. Yoon, *Data and Information Management Architecture for EOSDIS Core System*, submitted for publication.

[14] Menascé, D. A., L. Kerschberg, and H. Gomaa, *ECS Client-Server Systems Architecture*, The GMU ECS Federated Client-Server Architecture, Final Report, Chapter 2, Part III, George Mason University, August 31, 1994.

[15] Menascé, D. A., V. A. F. Almeida, and L. W. Dowdy, *Capacity Planning and Performance Modeling: from mainframes to client-server systems*, Prentice Hall, Englewood Cliffs, NJ, 1994.

[16] Menascé, D.A., S. C. S. Porto, and S. K. Tripathi, *Static Heuristic Processor Assignment in Heterogeneous Multiprocessors*, International Journal of High Speed Computing, Vol. 6, No. 1 (March 1994).

[17] Menascé, D. A. and S. Killedar, *Performance Modeling of ECS*, The GMU ECS Federated Client-Server Architecture, Final Report, Chapter 5, Part III, George Mason University, August 31, 1994.

[18] Menascé, D.A. and S. C.S. Porto, *Scheduling on Heterogeneous Message Passing Architectures*, Journal of Computer and Software Engineering, Ablex Publishing Co., New Jersey, Volume 1, Number 3, 1993.

[19] Menascé, D.A., D. Saha, S. C. da Silva Porto, V. A. F. Almeida, and S. K. Tripathi, *Static and Dynamic Processor Scheduling Disciplines in Heterogeneous Parallel Architectures*, accepted for publication in the Journal of Parallel and Distributed Computing.

[20] Menascé, D.A. and V.A.F. Almeida, *Cost-Performance Analysis of Heterogeneity in Supercomputer Architectures*, Proceedings of the ACM-IEEE Supercomputing'90 Conference, New York, NY, USA, November 12-16, 1990.

[21] Menascé, D.A. and V. A. F. Almeida, *Heterogeneity in High Performance Computing*, Proceedings of the 2^{nd} Symposium on High Performance Computing, October 7-9, Montpellier, France, 1991, published by Elsevier Science Publishers, eds: M. Durand and F. El Dabaghi.

[22] Menascé, D.A. and V. A. F. Almeida, *Heterogeneous Supercomputing: Why is it Cost-Effective?*, Supercomputing Review, August 1991, Vol. 4, No. 8.

[23] Waldspurger, C. A., T. Hogg, B. Huberman, J. O. Kephart, and W. S. Stornetta, *Spawn: A Distributed Computational Economy*, IEEE Transactions on Software Engineering, Vol. 18, No. 2, Feb. 1992.

[24] Krueger, P. and R. Chawla, *The Stealth Distributed Scheduler*, Proceedings of the 11^{th} Conference on Distributed Computing Systems, 1991.