EOS MLS Science Data Processing System: A Description of Architecture and Capabilities

David T. Cuddy, Mark D. Echeverri, Paul A. Wagner, Audrey T. Hanzel, and Ryan A. Fuller

***Abstract*—The Earth Observing System (EOS) Microwave Limb Sounder (MLS) is an atmospheric remote sensing experiment led by the Jet Propulsion Laboratory of the California Institute of Technology. The objectives of the EOS MLS are to learn more about stratospheric chemistry and causes of ozone changes, processes affecting climate variability, and pollution in the upper troposphere. The EOS MLS is one of four instruments on the National Aeronautics and Space Administration (NASA) EOS Aura spacecraft launched on July 15, 2004, with an operational period extending at least 5 years after launch. This paper describes the architecture and capabilities of the Science Data Processing System (SDPS) for the EOS MLS. The SDPS consists of two major components—the Science Computing Facility and the Science Investigator-led Processing System. The Science Computing Facility provides the facilities for the EOS MLS Science Team to perform the functions of scientific algorithm development, processing software development, quality control of data products, and scientific analyses. The Science Investigator-led Processing System processes and reprocesses the science data for the entire mission and delivers the data products to the Science Computing Facility and to the Goddard Space Flight Center Earth Science Distributed Active Archive Center, which archives and distributes the standard science products. The Science Investigator-led Processing System is developed and operated by Raytheon Information Technology and Scientific Services of Pasadena under contract with Jet Propulsion Laboratory.**

***Index Terms*—Computer facilities, data handling, data processing.**

I. Introduction

THE Earth Observing System (EOS) Microwave Limb Sounder (MLS), a passive microwave instrument [1], observes natural thermal radiation from the limb of the Earth's atmosphere. These observations yield the concentration at various heights of chemical species such as ozone and chlorine compounds and other atmospheric parameters such as temperature. EOS MLS makes global measurements, both day and night, that are reliable even in the presence of polar stratospheric clouds or volcanic aerosol [1], [2]. EOS MLS follows the very successful MLS on NASA's Upper Atmosphere Research Satellite [2] launched in 1991.

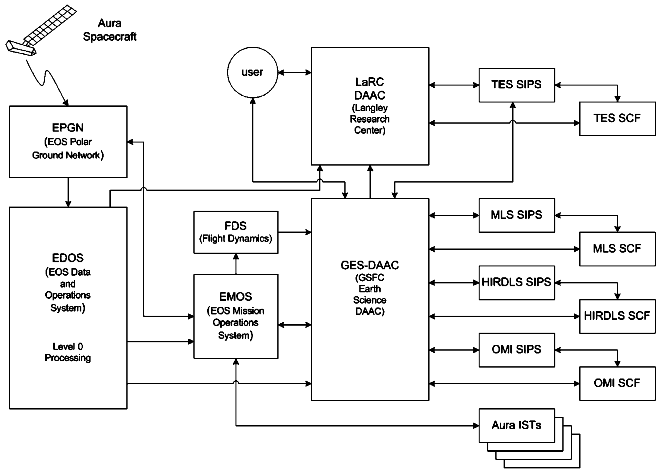
Aura Spacecraft
user
LaRC DAAC (Langley Research Center)
TES SIPS
TES SCF
EPGN (EOS Polar Ground Network)
EDOS (EOS Data and Operations System)
Level 0 Processing
FDS (Flight Dynamics)
EMOS (EOS Mission Operations System)
GES-DAAC (GSFC Earth Science DAAC)
MLS SIPS
MLS SCF
HIRDLS SIPS
HIRDLS SCF
OMI SIPS
OMI SCF
Aura ISTs

Fig. 1. Aura data flow architecture diagram.

The experiment is a result of collaboration between the United States and the United Kingdom, in particular the University of Edinburgh. The Jet Propulsion Laboratory (JPL), Pasadena, CA, has overall responsibility for instrument and algorithm development and implementation, along with scientific studies, while the University of Edinburgh Meteorology Department has responsibilities for aspects of data processing algorithm development, data validation, and scientific studies.

The MLS SDPS consists of two major components [3]—the JPL Science Computing Facility (SCF) and the Science Investigator-led Processing System (SIPS)—within a larger ground data system that was designed for the NASA EOS to support such missions as Terra, Aqua, and Aura. Except where explicitly stated otherwise, in this paper we focus on the US facilities. Other major components within the Aura ground data system, shown in Fig. 1, include EOS Polar Ground Network, EOS Data and Operations System (EDOS), Flight Dynamics, EOS Mission Operations System, the Goddard Space Flight Center (GSFC) Earth Science Distributed Active Archive Center (GES–DAAC), Langley Research Center DAAC, and users. The other instruments on Aura have science data processing systems similar to the MLS SDPS. The spacecraft data and instrument data flow to EDOS through the EOS Polar Ground Network with downlink stations in Alaska and Norway. EDOS is responsible for collecting the raw data, sorting it, time ordering it, removing redundancies, outputting the data in either production data sets (PDS) or as expedited data sets (EDS), and delivering the products to the appropriate DAAC for archive and distribution. EOS mission operations system (EMOS) responsibilities include the operations of the Aura spacecraft and the instruments and the processing of the Aura housekeeping data. The individual instrument teams

Manuscript received April 29, 2005; revised September 28, 2005. The work at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, was done under Contract NAS7-03001 with the National Aeronautics and Space Administration. The work at Raytheon was done under a contract with Jet Propulsion Laboratory.

D. T. Cuddy, P. A. Wagner, and R. A. Fuller are with the Jet Propulsion Laboratory, Pasadena, CA 91109 USA (e-mail: david.t.cuddy@jpl.nasa.gov).

M. D. Echeverri and A. T. Hanzel are with Raytheon Information Technology and Scientific Services, Pasadena, CA 91101 USA.

Digital Object Identifier 10.1109/TGRS.2005.863712

0196-2892/$20.00 © 2006 IEEE

TABLE I
INPUTS TO MLS SDPS. THE SHORT NAME IS USED AS THE HANDLE FOR EACH DATA TYPE WITHIN THE ECS ARCHITECTURE. THERE ARE SIX SEPARATE LEVEL 0 INSTRUMENT ENGINEERING DATASETS FOR EACH OF THE APIDs

Short Name	Collection Summary	Data Format	Daily size (MB)
ML0SCI1	MLS/Aura L0 Science Data APID=1744	CCSDS PDS	530.88
ML0SCI2	MLS/Aura L0 Science Data APID=1746	CCSDS PDS	530.88
ML0ENG1, 2, 3, 4, 5, 6	MLS/Aura L0 Instrument Engineering Packet 1 APID=1732, 1734, 1736, 1738, 1740, 1742	CCSDS PDS	36.0
AUREPHMH	Aura Satellite Definitive Ephemeris Data	HDF4	5.1
AURATTH	Aura Satellite Definitive Attitude Data	HDF4	5.208
D4FAEMIS	GMAO DAS First-look misc meteorology fields time averaged on eta coordinates	HDF-EOS	551.6
D4FAPMIS	GMAO tsyn3d_mis_p, DAS First-look 3d state (miscellaneous) instantaneous on pressure coordinates	HDF-EOS	180.2
D4FAXMIS	GMAO tsyn2d_mis_x, DAS First-look 2d (miscellaneous), instantaneous	HDF-EOS	31.7
D4LAEMIS	GMAO DAS Late-look misc meteorology fields time averaged on eta coordinates	HDF-EOS	551.6
D4LAPMIS	GMAO tsyn3d_mis_p, DAS Late-look 3d state (miscellaneous) instantaneous on pressure coordinates	HDF-EOS	180.2
D4LAXMIS	GMAO tsyn2d_mis_x, DAS Late-look 2d (miscellaneous), instantaneous	HDF-EOS	31.7
SAMOISTH	National Center for Environmental Prediction (NCEP) GDAS stratospheric analysis product – moisture/relative humidity	HDF-EOS	0.12
SATEMPH	National Center for Environmental Prediction (NCEP) GDAS stratospheric analysis product – temperature	HDF-EOS	0.44
SAWINDSH	National Center for Environmental Prediction (NCEP) GDAS stratospheric analysis product – U and V winds	HDF-EOS	0.55
SAGHGTH	National Center for Environmental Prediction (NCEP) GDAS stratospheric analysis product – geopotential height	HDF-EOS	0.54
AURGBAD1	1 Second GBAD Data (APID 967)	CCSDS PDS	19.2
LeapSecT	Leap Seconds file required for accurate SDP Toolkit coordinate system conversions	ASCII	0.01
UTCPoleT	Earth Motions file required for accurate SDO Toolkit coordinate system conversions	ASCII	0.01

work with EMOS using the EOS provided Instrument Support Terminals to monitor the health of the instruments and to provide commands to be up-linked to the spacecraft and the instruments. Flight Dynamics is responsible for the processing of the spacecraft orbit data.

There are two DAACs that provide the archive and distribution functions to the Aura mission and its four instruments. The other three companion instruments on Aura are the High Resolution Dynamics Limb Sounder (HIRDLS), the Ozone Monitoring Instrument (OMI), and the Tropospheric Emission Spectrometer (TES). The Langley Research Center DAAC provides support to the TES instrument, and the GES–DAAC provides support to OMI, HIRDLS, and MLS. In addition to supporting the spacecraft data and instrument data, GES–DAAC provides auxiliary data required for MLS science data processing, which are specified in Table I. MLS science software requires the earth motion data provided by the U.S. Naval Observatory, the meteorological data provided by the National Centers for Environmental Predictions (NCEP), and the meteorological data provided by the Global Modeling and Assimilation Office (GMAO). NCEP provides a set of combined stratospheric analysis products for temperature, humidity, geopotential height, and winds. GMAO provides both first look assimilation and late look assimilation products. The first look assimilation products use conventional and satellite observations available at the cutoff times to produce a timely set of atmospheric analysis within 6–10 h of the analysis times. The late look assimilation products use a software configuration that is identical to the first look products but use a more complete set of input observations and are produced after a delay of about two weeks. The GES–DAAC is also responsible for the archive and distribution of the standard data products produced by the MLS SDPS.

TABLE II
SUMMARY OF DATA VOLUMES FOR THE MLS STANDARD PRODUCTS. THE VOLUME NUMBERS DO NOT INCLUDE ENGINEERING, DIAGNOSTICS, CALIBRATION, AND LOG FILES THAT ARE GENERATED IN THE PROCESS OF GENERATING THE STANDARD PRODUCTS

Data Sets	Volume (MB)	Granule Count	Yearly Volume (GB)
Level 0	1,097/day	96/day	400
Level 1	4,142/day	4/day	1,512
Level 2	862/day	21/day	315
Level 3 daily map	52/day	13/day	19
Level 3 daily zonal mean	37/day	2/day	13
Level 3 monthly	115/month	4/month	1.4
Other data	1,558/day	37/day	569
Total	**7,748/day & 115/month**	**171/day & 4/month**	**2,829**

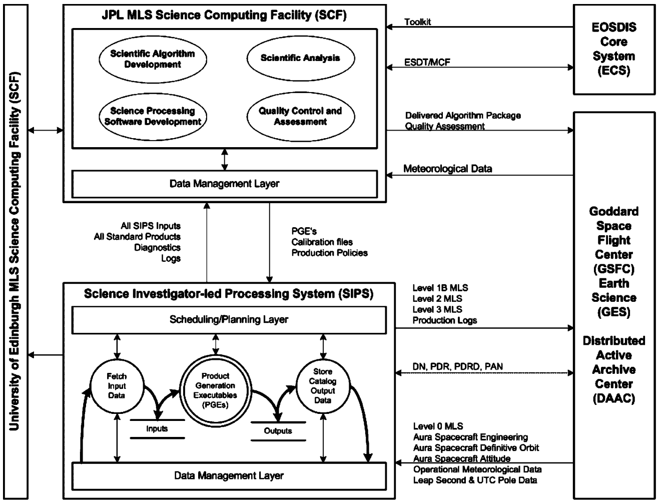
University of Edinburgh MLS Science Computing Facility (SCF)
JPL MLS Science Computing Facility (SCF)
Scientific Algorithm Development
Scientific Analysis
Science Processing Software Development
Quality Control and Assessment
Data Management Layer
All SIPS Inputs
All Standard Products
Diagnostics
Logs
PGE's
Calibration files
Production Policies
Science Investigator-led Processing System (SIPS)
Scheduling/Planning Layer
Fetch Input Data
Product Generation Executables (PGEs)
Store Catalog Output Data
Inputs
Outputs
Data Management Layer
Toolkit
ESDT/MCF
Delivered Algorithm Package
Quality Assessment
Meteorological Data
Level 1B MLS
Level 2 MLS
Level 3 MLS
Production Logs
DN, PDR, PDRD, PAN
Level 0 MLS
Aura Spacecraft Engineering
Aura Spacecraft Definitive Orbit
Aura Spacecraft Attitude
Operational Meteorological Data
Leap Second & UTC Pole Data
EOSDIS Core System (ECS)
Goddard Space Flight Center (GSFC) Earth Science (GES)
Distributed Active Archive Center (DAAC)

Fig. 2. MLS SDPS Context Diagram.

II. SCIENCE DATA PROCESSING SYSTEM

The main function of SDPS is to produce higher level science data products for EOS MLS. Table II gives the data volumes for MLS data by collection sets. The context diagram for SDPS is shown in Fig. 2. The SDPS performs this function using two major subsystems—SCF and SIPS.

The SCF provides a system of resources to the Science Team for scientific analyses, algorithm development, science software development, data quality control and assessment, and special data production. The SCF includes a data management layer that accepts and stores the incoming data products for access by the Science Team. The U.K. SCF has its own separate facility and provides the same services as their US colleagues.

Raytheon Information Technology and Scientific Services developed the SIPS under contract with JPL, and they operate the system around the clock but provide personnel only during prime shift. The SIPS provides a system to produce the standard science data products through processing and reprocessing using algorithms provided by the MLS science team. The SIPS controls data flow and stores data using a data management layer and provides control to the operator using a scheduling/planning layer.

III. Interfaces

A. Interface Between GES–DAAC and SIPS

The GES–DAAC provides spacecraft data, instrument data, earth motion data, and meteorological data [4] to the SIPS as these data become available using the subscription mechanism. Table I lists the products that are sent from GES–DAAC to the SIPS. The PDS are provided in uniform 2-h segments, 12 times per day. The products are pushed to a secured copy server at the SIPS over the EOS provided network. Once the transfer is complete, GES–DAAC sends a Distribution Notification via an e-mail. The full details of this protocol are described in the Interface Control Document between the ECS and SIPS [5]. Upon receiving an e-mail for the Distribution Notification, the SIPS ingests the products into its system and removes the products from the secure copy server.

The SIPS provides its higher level products to the GES–DAAC using a product delivery record (PDR) mechanism that uses a secure copy server at the SIPS. The SIPS posts the products in a disk directory and a related PDR in a pre-agreed directory. The GES–DAAC polls this pre-agreed directory for new PDRs and when found uses the information in the PDR to retrieve the products from the directory specified therein. Once the GES–DAAC has retrieved the products and has successfully archived the products, it sends a Product Acceptance Notice to the SIPS via e-mail. The SIPS then removes the product from the secure copy server. The SIPS uses the Machine-to-Machine Gateway [6] to check once per day to assure that the contents of its own data holdings match the data holdings at the GES–DAAC. If they do not match, either a request is placed with the GES–DAAC to retrieve the missing product, or a subscription order is placed in the SIPS to redeliver the products missing in the GES–DAAC archives.

B. Interface Between GES–DAAC and SCF

The GES–DAAC provides the SCF with the EDS products and the GMAO meteorological data using the very same subscription mechanism used to deliver products to the SIPS, except the secure copy server in this case is provided by the SCF. The SCF ingests the incoming products and removes the data from the secure copy server. The EDS products are provided only on request and differ from PDS in two respects. The time coverage is based on satellite contact period rather than the uniform two hour period, and the data is provided on an expedited basis. The GMAO products received at the SCF are the first-look products that are used for regular and timely inspections of MLS products and the late look products that are used for research.

The SCF provides the GES–DAAC with the delivered algorithm package (DAP) and the associated quality documents with each new version of the product generation executables (PGEs) used at the SIPS to generate higher level products. These new versions of the DAP and corresponding quality documents occur very infrequently, and the Science Team manually provides them to the GES–DAAC operations.

C. Interface Between SIPS and SCF

The SIPS provides data to the SCF using the very same PDR mechanism used with the GES–DAAC with a slight modification. Once it successfully obtains the products, the SCF deletes the PDR to signal the success to the SIPS rather than sending a Product Acceptance Notice by e-mail. The SIPS sends all data including all inputs from GES–DAAC, all higher level science products, and associated engineering, diagnostic, and log files to the SCF.

Because the limited bandwidth (200 Kb/s) from the U.S. to U.K. does not justify sending all data via secure copy and because the U.K. SCF does not have adequate online storage, the SIPS operations staff copies all data to DVD media, which it sends via regular mail on a periodic basis to the MLS co-investigators at the University of Edinburgh. The SIPS operations staff copies a limited set of data to DVD for the SCF for safekeeping in case of disk outages.

The Science Team at the SCF provides the SIPS with the PGEs and associated configuration and processing files for each version of the PGE in the form of a DAP. This action is taken with careful oversight and under strict configuration management. The DAP includes source code, a description of the processing methodology, test data, a description of the data products, required metadata, and executables for each PGE.

IV. Science Computing Facility

The SCF provides the services and resources to the EOS MLS Science Team to perform scientific algorithm development, science processing software development, scientific quality control, and scientific analysis. The SCF provides a distributed network of computer systems with high-performance computers and large file servers for use by the Science Team. The Science Team uses the SCF to develop, run, and test the PGEs, to produce any special products, and to perform scientific analyses, algorithm development, and data validation.

In order to support the development of the PGEs, the SCF has very similar processing systems to the SIPS. The SCF provides additional processors to support the scientific analyses, data validation, and data quality control. The SCF employs computing clusters to provide the required processing power. At the time of this writing, the total number of nodes in the nonhomogeneous SCF clusters is approximately 500 with a Composite Theoretical Performance[1] value of about 5 trillion theoretical operations per second.[2] To support the large storage requirement, the SCF employs a network file system that currently has about 8 TB of online storage capable of growing to many more terabytes. The SCF employs a tape robotic system with multiple tape drives to provide backup storage of the online storage. All data that can be easily reproduced are not put to backup storage. All backup storage also has an off-site storage to aid recovery from localized disaster. The SCF provides plotting capability with plotters and color printers so that the Science Team can visualize the data quality graphically.

[1] http://www.access.gpo.gov/bis/ear/pdf/ccl4.pdf.

[2] http://support.intel.com/support/processors/sb/ CS-017 346.htm.

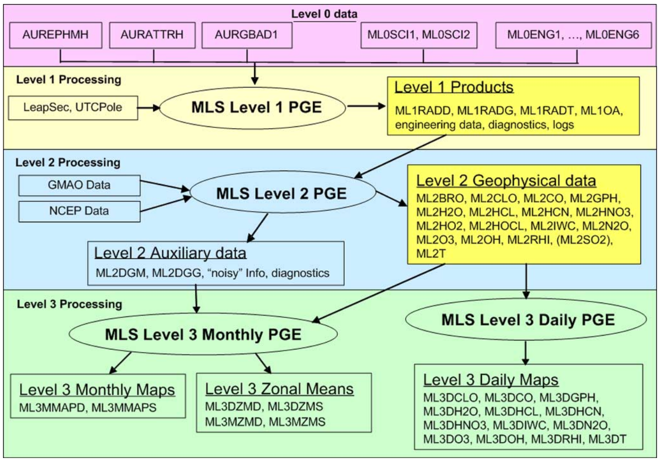

Fig. 3. EOS MLS science data flow diagram. ML2SO2 is produced only when volcanic activities generate sufficient particles in the upper atmosphere. Lines LeapSec, UTCPole box to the MLS Level 3 Monthly and Daily PGEs were not drawn only to avoid clutter, but these files are used by these PGEs as well.

To manage the very large storage system, the SCF arranges its directories in hierarchical layers using the data source, data type, processing version, data observation year and date. All data from EOS MLS are found beneath one master directory, and in that directory each data type has its own subdirectory. In each of these data type subdirectories, there are further subdirectories for the processing version of the producing PGE. The data is further organized by data observation year and day of year. In some cases, a directory for the day of year may not be used if only one product per day is produced. The rule of thumb guiding this layering and organizing is to limit the number of files in any given directory to less than one thousand.

Each product usually has the data file and an associated metadata file that contains the descriptive information required to identify the data. The description includes identity, production date and time, time coverage, quality flags and descriptions, geographical extent, processor identity and version. MLS together with the other three instruments on Aura chose to use similar file formats and naming schemes [7] in which each granule is given a unique name based on instrument, spacecraft, data type and subtype, processor version, cycle number, data time, and data format. The data kept in the SCF are also catalogued in a database so that data access can be optimized, organized, and linked with other information such as data plots, science analysis information, instrument behavior, and data quality assessments.

V. Product Generation Executables

The PGEs process the incoming Level 0 data to Level 1B, Level 2, and Level 3 data products, successively. The PGEs may be executed independently at the SCF or within the SIPS framework. Fig. 3 shows the data flow amongst the PGEs. The Science Data Processing Toolkit that is supplied by Earth Science Data and Information System Project provides a utility layer for the PGEs. To accomplish this, the Toolkit provides a common set of routines to handle inputs and outputs, messaging, error handling, time, spacecraft geometry, planetary orbits, and instrument geometry. In each PGE, the Toolkit requires a process control file that provides a mechanism for identifying all input files, all output files, and run-time processing parameters. Additionally, MLS employs a configuration file for each PGE that determines the behavior of the PGE during execution. The configuration files use a functional processing mini-language that allows the user to specify data flow, commands, parameters, and declarations. This behavior is an essential part of the algorithms. For data production at the SIPS, each of these files remains static, however at the SCF each run may employ a different configuration file, thereby allowing the same executable to behave in a different way with the same input files.

TABLE III
MLS Level 1B Standard Products. All of These Use the HDF5 Format

Short Name	Description	Daily Size (MB)
ML1BOA	Level 1B Orbit and Attitude	306
ML1BRADD	Level 1B Radiances for the DACS	1,853
ML1BRADG	Level 1B Radiances for the GHz	1,528
ML1BRADT	Level 1B Radiances for the THz	455

In order to make software code easier to read and easier to maintain, MLS developed programming guidelines [8] to be used in the production code. MLS chose to use Fortran 95 to implement the PGEs and established guidelines to restrict how this language is used. The PGEs do not use some features of the language including the Fortran 77 statements that have become obsolete and those that are destined to become obsolete in future Fortran standards. MLS restricts the use of Fortran-provided input and output statements in production code; instead MLS relies on appropriate procedures provided in libraries such as the Toolkit, HDF, and HDF-EOS packages. MLS further restricts coding practices by using naming conventions for keywords, intrinsic functions and subroutines, constants, variables, and modules. MLS employs a message layer that handles four levels of severity, which are debug, info, warning, and error. MLS uses a set of programming styles and coding standards to establish consistency of software modules and enhance maintenance. All PGEs execute in the context of a script that operates under the Linux operating system with the IA32 architecture.

The Level 1 Processor accepts the Level 0 input (instrument data counts—science and engineering) and the spacecraft ancillary data, and it produces the Level 1B product (calibrated radiances) as the main product. The Level 0 science and engineering data arrive in granularity of 2 h; however the Level 1 Processor produces Level 1B outputs in granularities of a day. It also produces associated engineering and diagnostic data. The outputs of the Level 1 Processor are shown in Table III. The reader should refer to the paper on the Level 1 algorithm [9] for more details about this PGE. To process a full day's data, the Level 1 Processor requires less than 6 h on a 3-GHz Intel Xeon processor with at least 2 GB of memory.

The Level 2 Processor accepts the Level 1B products and operational meteorological data and produces a set of Level 2 products (geophysical parameters at full resolution). It also produces diagnostic information, ancillary data, and summary logs. The outputs of the Level 2 Processor are shown in Table IV. The reader should refer to the paper on the Level 2 algorithm [10] for more details about this PGE. The Level 2 Processor requires significant computational resources. In order to process

TABLE IV
MLS Level 2 Geophysical Products. All Products Use the HDF-EOS5 Swath Except ML2DGM, Which Uses the Plain HDF5 Format

Short Name	Description	Daily Size (MB)
ML2BRO	L2 Bromine Monoxide (BRO) Mixing Ratio	2.57
ML2CLO	L2 Chlorine Monoxide (CLO) Mixing Ratio	2.57
ML2CO	L2 Carbon Monoxide (CO) Mixing Ratio	2.57
ML2DGG	L2 Diagnostics, Geophysical Parameter Grid	217.5
ML2DGM	L2 Diagnostics, Miscellaneous Grid	597.7
ML2GPH	L2 Geopotential Height	2.17
ML2H2O	L2 Water Vapor (H2O) Mixing Ratio	2.56
ML2HCL	L2 Hydrogen Chloride (HCL) Mixing Ratio	2.57
ML2HCN	L2 Hydrogen Cyanide (HCN) Mixing Ratio	2.56
ML2HNO3	L2 Nitric Acid (HNO3) Mixing Ratio	2.56
ML2HO2	L2 Hydroperoxy (HO2) Mixing Ratio	2.56
ML2HOCL	L2 Hypochlorous Acid (HOCL) Mixing Ratio	2.56
ML2IWC	L2 Ice with Respect to Cloud Product	2.97
ML2N2O	L2 Nitrous Oxide (N2O) Mixing Ratio	2.56
ML2O3	L2 Ozone (O3) Mixing Ratio	2.56
ML2OH	L2 Hydroxyl (OH) Mixing Ratio	2.56
ML2RHI	L2 Relative Humidity With Respect To Ice	2.17
ML2SO2	L2 Sulfur Dioxide (SO2) Mixing Ratio	2.56
ML2T	L2 Temperature	3.14

one data day, the Level 2 Processor requires between 20 and 30 h on 350 Intel Xeon processors clocked at 3 GHz. MLS employs a cluster of processors connected by a gigabit Ethernet. The Level 2 Processor splits one day of Level 1 data into 350 chunks and sends these 350 chunks to 350 separate processors. After all 350 processors complete their processing, the outputs from them are sewn together into outputs with granularities of a day. If there are fewer than 350 processors, additional cycles of processors are required after the first round of chunks are completed. If the Level 2 Processor is to finish a data day in one cycle, it requires a minimum of 350 processors. At launch the SIPS configured a cluster with 364 Intel Xeon processors. The extra 14 gave a 4% margin to account for possible computer outages. This system allows the SIPS to process five data days each week skipping the remaining two days, which meets the requirements to process 60% of Level 2 for which it was funded and designed for the first year of processing. Additional capability is now being added that will double the throughput.

In order to maximize the use of any number of processors, a feature of the Level 2 Processor called the Queue Manager coordinates the use of the processors by requests from the master jobs. The master job manages the chunks for each day, and for each chunk the master job requests the exclusive use of a processor from the Queue Manager. The Queue Manager allocates a free processor to the master job and marks the processor as "in use" preventing other master jobs from using that processor. Once the slave job for the chunk has completed, the master job releases the processor back to the Queue Manager, and the Queue Manager puts that processor back on the list of available processors. Studies have shown that we can gain up to 30% efficiency if the number of processors exceeds the number of chunks in a day by taking advantage of idle processors that finish before the slowest chunk in the day.

TABLE V
MLS Level 3 Daily Map Products. All Products Use the HDF-EOS5 Grid Format

Short Name	Description	Daily Size (MB)
ML3DCLO	L3 Daily Map of Chlorine Monoxide (CLO) Mixing Ratio	3.71
ML3DCO	L3 Daily Map of Carbon Monoxide (CO) Mixing Ratio	5.99
ML3DGPH	L3 daily map of Geopotential Height	4.93
ML3DH2O	L3 Daily Map of Water Vapor (H2O) Mixing Ratio	4.93
ML3DHCL	L3 Daily Map of Hydrogen Chloride (HCL) Mixing Ratio	3.17
ML3DHCN	L3 Daily Map of Hydrogen Cyanide (HCN) Mixing Ratio	1.06
ML3DHNO3	L3 Daily Map of Nitric Acid (HNO3) Mixing Ratio	2.12
ML3DIWC	L3 Daily Map of Cloud Ice Product	3.17
ML3DN2O	L3 Daily Map of Nitrous Oxide (N2O) Mixing Ratio	2.47
ML3DO3	L3 Daily Map of Ozone (O3) Mixing Ratio	8.46
ML3DOH	L3 Daily Map of Hydroxyl (OH) Mixing Ratio	4.23
ML3DRHI	L3 Daily Map of Relative Humidity With Respect To Ice	3.17
ML3DT	L3 Daily Map of Temperature	4.93

TABLE VI
MLS Level 3 Monthly Products. The L3 Daily Zonal Means Have the Granularity of a Day, However They Are Produced by the MLS Level Three Monthly PGEs. The Zonal Mean Products Use the HDF-EOS5 Zonal Mean Format and the Month Maps Use the HDF-EOS5 Grids

Short Name	Description	Monthly Size (MB)
ML3DZMS	L3 Daily Zonal Means, Standard Products	12.3
ML3DZMD	L3 Daily Zonal Means, Diagnostic Products	24.6
ML3MMAPD	L3 Monthly Maps, Diagnostic Products	70.95
ML3MMAPS	L3 Monthly Maps, Standard Products	43.23
ML3MZMS	L3 Monthly Zonal Means, Standard Products	0.49
ML3MZMD	L3 Monthly Zonal Means, Diagnostic Products	0.82

With the next release of Level 2 software, reprocessing will require a great amount of the resources of the MLS system. If the next release were to be made two years into the mission, it will take about one year to reprocess the backlog of data in addition to keeping up with the incoming data. This can be accomplished by utilizing the two SIPS clusters and one of the SCF clusters that will be temporarily attached to the SIPS. With the Queue Manager in place on all three systems, we expect to be able to process 21 days of data for every week. In the current design of data flow, we expect to use first look GMAO products in the reprocessing.

The Level 3 Processor consists of two PGEs—Level 3 Daily and Level 3 Monthly. The Level 3 Daily accepts a set (equivalent to 30 days starting from the time of instrument activation) of standard Level 2 products (produced by the Level 2 Processor) and produces a set of Level 3 products in the form of gridded daily maps. The outputs of Level 3 Daily are shown in Table V. Level 3 Monthly accepts a set of standard Level 2 products and a set of Level 2 auxiliary data products, and it produces a set of daily zonal means, gridded monthly average maps, and monthly zonal means for each calendar month. The outputs of Level 3 Monthly are shown in Table VI. The reader should refer to the paper on the Level 3 algorithm [11] for more details about these two PGEs.

VI. Science Investigator-Led Processing System

The SIPS provides a production system for EOS MLS to produce standard science data products. The SIPS provides the con-

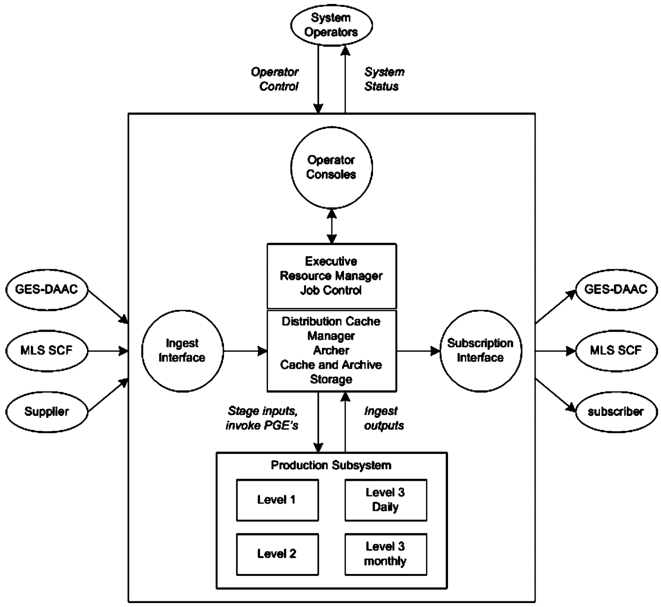

Fig. 4. MLS SIPS architecture diagram. "Supplier" and "Subscriber" show how other possible suppliers and subscribers can be easily plugged into this architecture.

trol and data management of the inputs and outputs and the environment for the execution of the PGEs. Fig. 4 diagrams the SIPS architecture. The SIPS interfaces with GSFC-DAAC to receive EOS MLS Instrument Level 0 Science and Engineering data, Aura Spacecraft Engineering data, and Operational Meteorological Data. The SIPS delivers the standard data products shown in Tables II–VI to GES–DAAC for archive and distribution. The SIPS delivers all input data plus the standard data products, diagnostics, and log files to the SCF for use and validation by the Science Team. The SIPS receives the DAP, the production control and configuration files, and the processing policies from the SCF that are used in production.

The SIPS makes extensive reuse of design and code [12] from the Vegetation Canopy LIDAR Data Center (VDC) which in turn evolved from the V0 that was developed in the 1990s for the GSFC DAAC. Because much of it is inherited, the software used in the SIPS is mostly in C and C++ using SQL calls to a relational database. The SIPS operates on Sun computers using the Solaris operating system and Korn shell scripts. It interfaces with other platforms running a version of the Linux operating system that host the PGEs.

The SIPS is a production data system, and as in any well controlled production system there is detailed tracking of inputs, outputs, and production engines. The SIPS is designed for high-volume, high-density data and is batch oriented.

The SIPS employs a relational database to inventory the information about data as they are received, stored, created, processed, and distributed. The tracking attributes include file version, data start and end times within the file, EOS metadata attributes, identity, time of action, type of action, locations, versions, volume, originator, destination, and data type.

The SIPS uses a message passing layer [13] to enable various system components to communicate with each other. This layer allows any system component to act as a server or a client or to engage in a peer-to-peer communications. It facilitates the SIPS as a distributed system to run on many hosts. The message passing design allows flexibility in message definitions and easy transmission of complex data structures. The message passing can be either one way (notification) or two ways (request/response).

All work in the SIPS occurs in the context of "jobs" managed by a batch manager subsystem called the executive. A job is a collection of processes that accomplishes a task. The executive monitors the execution of each step in the job and if a step fails, the job is considered to have failed. There are three types of jobs: ingest, science, and distribution. An ingest job places the granule under the ownership of the SIPS by identifying, cataloging and storing the data granule. A science job invokes executable modules to generate data products. All science jobs fetch inputs, execute a PGE, and store outputs. Note that the store action triggers one or more ingest jobs for the newly created products. The PGEs run on a different set of hosts than the SIPS hosts and return either a success or a failure at the end of the execution. A distribution job runs to stage the SIPS generated products for external interfaces. The primary external interface is a file server that allows trusted hosts to retrieve the products using the PDR mechanism.

The resource manager subsystem acts as an accountant for the resources within the SIPS. There are three types of resources: disk partitions, work directories, and discrete resources. Resources are requested and granted on an all-or-nothing basis to minimize dead-lock conditions.

The job scheduler subsystem allows auto-planning based on a set of work flow rules that include required inputs, data availability timeouts, and PGE version. The job scheduler also allows manual planning by an operator.

The SIPS provides a large amount of storage (terabytes) including the use of tapes and CDs or any device whose driver allows access through UNIX's logical file system. The SIPS uses a collection of system components for managing the large storage. These components include a monitor, gateway service, get/put functions, media manager, and library manager.

VII. Conclusion

The SDPS for EOS MLS met all science data processing requirements by assuring the effective cooperation of its components widely dispersed in location and under the responsibility of different institutions. Each component exercises control over its operations and exchanges data as needed with other components by reliable mechanisms. This accomplishes several design goals. Allowing decisions to be made at a local level permits the operator to maximize that component's performance. Well-defined interfaces guarantee robustness of the SDPS as a whole. Finally, any problems that may occur are easily localized, diagnosed, and corrected.

Acknowledgment

The authors would like to thank J. Waters, D. Flower, and N. Livesey for their support and valuable comments they have provided. They also thank colleagues at JPL, especially R. Jarnot, V. Perun, V. Snyder, R.Thurstans and N. Patel for their help throughout the development and testing of the EOS

MLS Science Data Processing System. They thank the support they received from the GSFC ESDIS office, especially H. K. Ramapriyan, G. Iona, S. Scott, T. Goff, and K. Michael. They thank the Raytheon ITSS team, especially E. Greene, C. Vuu, and the system administration group.

REFERENCES

[1] J. W. Waters *et al.*, “The Earth Observing System Microwave Limb Sounder (EOS MLS) on the Aura satellite,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 5, pp. 1075–1092, May 2006.

[2] ——, “The UARS and EOS Microwave Limb Sounder experiments,” *J. Atmos. Sci.*, vol. 56, pp. 194–218, 1999.

[3] D. Cuddy, Functional requirements and design of the EOS MLS science data processing system, 2004.

[4] *Interface Control Document Between the EOSDIS Core System (ECS) and the Science Investigator-Led Processing System (SIPS) Volume 8 Microwave Limb Sounder (MLS) ECS Data Flows*, Dec. 2004. GSFC 423-41-57-8, Rev. B.

[5] *Interface Control Document between the EOSDIS Core System (ECS) and the Science Investigator-Led Processing Systems (SIPS) Volume 0 Interface Mechanisms*, Oct. 2002. GSFC 423-41-57-0, Revision F.

[6] *Interface Control Document Between the EOSDIS Core System (ECS) and the Science Investigator-Led Processing Systems (SIPS) Volume 9 Machine-to-Machine Search and Order Gateway*, Sep. 2002. GSFC 423-41-57-9, Revision A.

[7] C. Craig, K. Stone, D. Cuddy, S. Lewicki, P. Veefkind, P. Leonard, A. Fleig, and P. Wagner, “HDF-EOS Aura file format guidelines,” NCAR, Boulder, CO, NCAR Doc. SW-NCA-079, Ver. 1.3, 2003.

[8] D. Cuddy, “EOS MLS science data processing software development standards guide,” JPL, Pasadena, CA, Doc., Ver. 1.3, 2000.

[9] R. F. Jarnot, H. M. Pickett, and M. J. Schwartz, “EOS MLS Level 1 data processing algorithm theoretical basis,” JPL, Pasadena, CA, Doc. D15210, Ver. 2.0, 2004.

[10] N. J. Livesey and W. V. Snyder, “Retrieval algorithms for the EOS Microwave Limb Sounder,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 5, pp. 1144–1155, May 2006.

[11] Y. Jiang, “EOS MLS Level 3 algorithm theoretical basis,” JPL, Pasadena, CA, Doc. D-18911, 2005.

[12] M. Echeverri and A. Griffin, “Software design and reuse for a low-cost data processing infrastructure,” in *Proc. IGARSS*, vol. 3, 2001, pp. 1462–1464.

[13] M. Echeverri, “An overview of aura MLS data processing,” *Proc. SPIE*, vol. 4483, pp. 301–309, 2001.

David T. Cuddy received the B.A degree from the University of Oregon, Eugene, in 1974 and the M.S. degree from the University of Hawaii, Manoa, in 1976. His studies were in information and computer science.

He served in the U.S. Army from 1970 to 1973. He was with the Research Corporation of the University of Hawaii from 1976 to 1985 and was responsible for the Shipboard Computer Facility for the University. Since 1985, he has been with the Jet Propulsion Laboratory, Pasadena, CA, where he has worked on the NASA Scatterometer project and on the Alaska SAR Facility Development project before joining the MLS project in 1999. He currently manages the science software development and the science data production for the MLS.

Mark D. Echeverri received the B.S. degree in aerospace engineering from the University of California, Los Angeles, in 1993.

In 1999, he joined Raytheon ITSS, Pasadena, CA, working on the MLS SIPS as the Lead System Engineer.

Paul A. Wagner received the B.S. degree in physics from the California Institute of Technology, Pasadena, in 1976.

He has been with the Jet Propulsion Laboratory, Pasadena, since 1979. He is currently the Lead Software Engineer for the Level 2 production software.

Mr. Wagner is a member of the Acoustical Society of America.

Audrey T. Hanzel received the B.S. degree in mathematics and computer science from the University of California, Los Angeles, in 1983 and the M.S. degree in computer science from Loyola Marymount University, Los Angeles, in 1988.

She was a Software Engineer at Xerox Corporation, El Segundo, CA, from 1984 to 2000. She is currently the MLS SIPS Operations Manager and has been supporting the EOS MLS SDPS and SIPS as the Test Lead since joining Raytheon ITSS, Pasadena, CA, in 2000.

Ryan A. Fuller received the B.S. degree in computer science from the University of Colorado, Boulder, in 2002.

He has been with Jet Propulsion Laboratory, Pasadena, CA, since 2002.