

Grids for Dummies

Featuring Earth Science Data

Mining Application

Thomas H. Hinke

NASA Ames Research Center
Moffett Field, California, USA



Ames Research Center

1



Division

Outline

Use of Grids for Applications

- What are grids
- Grids from a user's perspective
- Grid support for Earth Science applications such as data Mining

Global Grid Forum

- Background
- Organization
- Current work

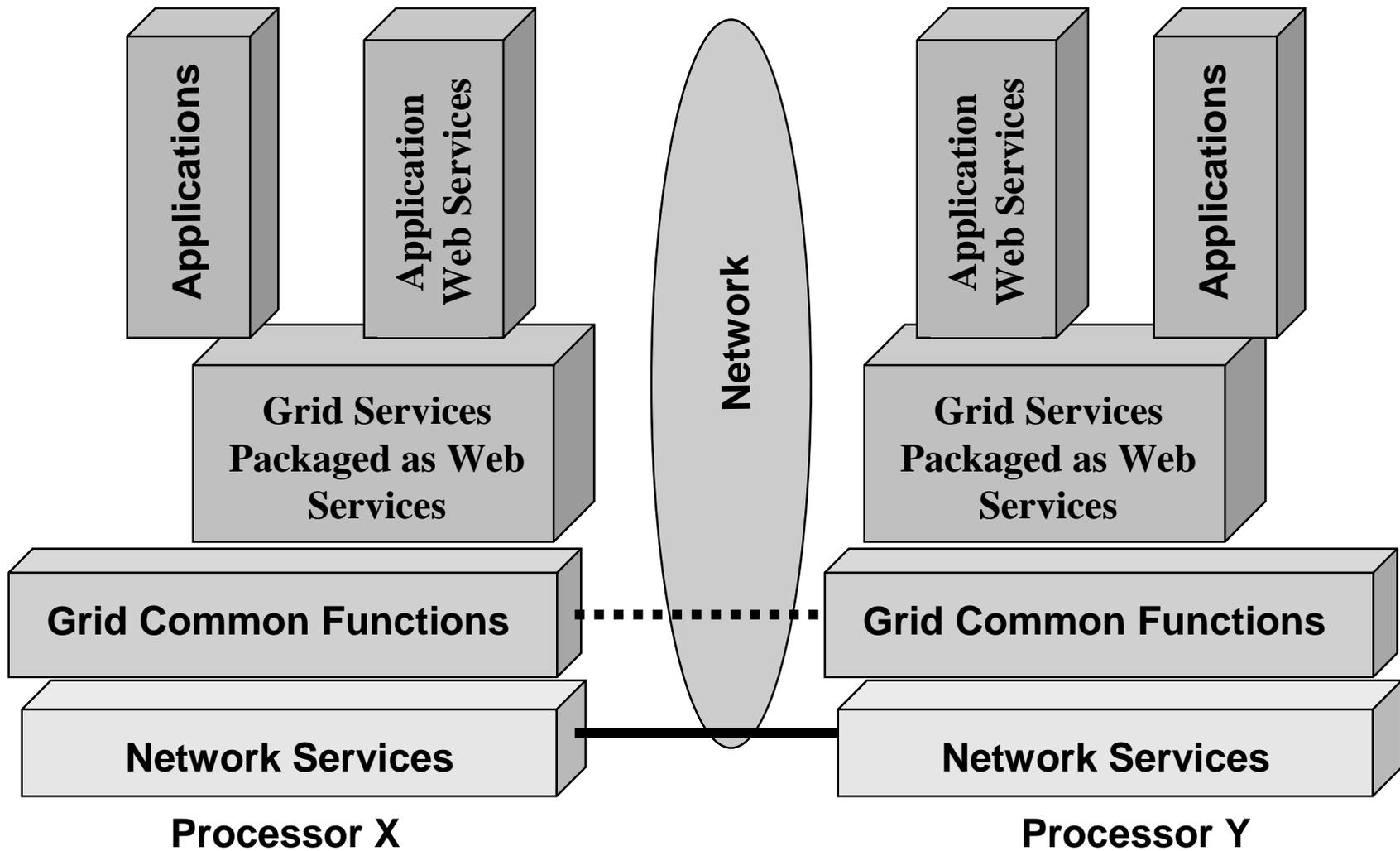


What Are Grids?

“Grids are persistent environments that enable software applications to integrate instruments, displays, computational and information resources that are managed by diverse organizations in widespread locations.” [<http://www.globus.org/>]



Middleware Makes the Grid



Characteristics Usually Found in Grids

An underlying security infrastructure such as the Grid Security Infrastructure (GSI), which is based on public key technology

- Protection for at least authentication information as it flows from resource to resource

Single sign-on

A seamless processing environment

An infrastructure that is scalable to a large number of resources

The ability for the grid components to cross administrative boundaries



Why are Grids Important?

Computing and data Grids are emerging as the infrastructure for 21st century science, engineering and high-performance applications and systems

- Grids provide a common way of managing distributed computing, data, instrument, and human resources

Grids facilitate collaboration by providing the glue of large-scale science and engineering.

- A common way to access and use shared data and simulations
- A common security model to facilitate the interaction of many different people from many different institutions

Grids provide a middle-ware environment that eases the development of complex systems.

- Grids can facilitate the development of large-scale science, engineering and operational applications
 - That are widely distributed
 - That are processing and/or data intensive



How the User Sees a Grid

A set of grid functions that are available as

- Application programmer interfaces (APIs)
- Command-line functions

After authentication, functions can be used to

- Spawn jobs on different processors with a single command
- Access data on remote systems
- Move data from one processor to another
- Support the communication between programs executing on different processors
- Discover the properties of computational resources available on the grid using the grid information service
- Use a broker to select the best place for a job to run and then negotiate the reservation and execution (coming soon).



What Will Grids Provide?

Support for collaboration

- Common authentication and security infrastructure
- Common mechanisms to share data
- Common mechanisms to access computing resources
- Management of community databases

Uniform data access

- Standardized mechanisms for accessing archival datasets
- Common mechanisms for managing metadata

Support for building systems

- Very few applications use a single computer
- At least some of the resources needed to solve one's problem invariably reside elsewhere
- Grids will supply the core capabilities common to most applications, so that application developers do not have to re-implement this core capability with each application



Web Access to the Grid is Available

Some web portals exist for accessing grids

- LaunchPad
 - Developed as part of the NASA Information Power Grid project
 - Uses Java Server Pages and Java Beans
 - Built using the Grid Portal Development Kit
- GridPort
 - Developed at the San Diego Super Computer Center
 - Uses Perl



How an Application Developer Sees a Grid

A set of grid functions

A set of grid functions packaged as web services

- Interface is defined through WSDL (Web Services Description Language)
- Standard access protocol is SOAP (Simple Object Access Protocol)



What a User Gains By Using a Grid

As a direct user

– Can easily

- Execute jobs at one or more remote sites
- Move data between sites
- All with single sign-on security

As a user of a grid enabled application

- Will not see the grid
- Will see an application whose development was eased with grid functions or grid-based web services
- Ease of development should result in more applications or faster availability of applications



What Application Developers Gain by Using Grids

Application web services can be built by re-using capabilities provided by existing grid-enabled Web services.

Applications can also be built by using grid functions

Grid functions/services handle distributed management of tasks and data

- Developer can focus on logic of application and not logic of distributed interaction



Grids Support Various Communities of Use

Scientists and domain problem solvers and other users

- They will use the applications and services that the grid facilitates.
 - They need to be able to express a problem or experiment in application domain-specific terms, specify the drivers (initial conditions, live data sources, etc.) request that the solution be obtained, and manage the resulting graphics, data, etc.

Model builders and computational scientists

- They will use the grid directly to realize their models and simulations.
 - They combine knowledge of the real world with theoretic models of the real world to produce simulations or models that can produce a “complete” representation of the observables

Application developers

- They will use the grid directly to realize applications that require high performance computing or a large number of distributed processors.
- They will use the models and simulations as components

Service builders

- They will build the frameworks that allow application developers to
 - Build grid services that can be used directly or
 - Use services as building blocks to more easily develop more complex services targeting specific application areas.



Summary of What User Gains

User can focus on solving domain issues of the problem and not on computer science issues of distributed computing



Most Grids Are Built on the Globus Toolkit

NASA's Information Power Grid (IPG) is one such example

The Globus project involves research and development personnel from

- Argonne National Laboratory
 - University of Southern California's Information Sciences Institute
 - NASA's Ames Information Power Grid Team
 - National Science Foundation PACI (Partnerships for Advanced Computational Infrastructure) programs at
 - National Center for Supercomputing Applications (NCSA)
 - San Diego Supercomputer Center
 - A number of universities
-



Data Mining on the Grid

What is data mining?

Why mine on the Grid?

The Grid Miner developed for NASA's
Information Power Grid (IPG)

A proposed IPG Mining Service



What Is Data Mining

“Data mining is the process by which information and knowledge are extracted from a potentially large volume of data using techniques that go beyond a simple search through the data.” [NASA Workshop on Issues in the Application of Data Mining to Scientific Data, Oct 1999, http://www.cs.uah.edu/NASA_Mining/]



Grid Miner

Developed as one of the early applications on the IPG

- Helped debug the IPG
- Provided basis for satisfying one of two major IPG milestones last year

Provides basis for what could be an on-going Grid Mining Service



Example: Mining for Mesoscale Convective Systems

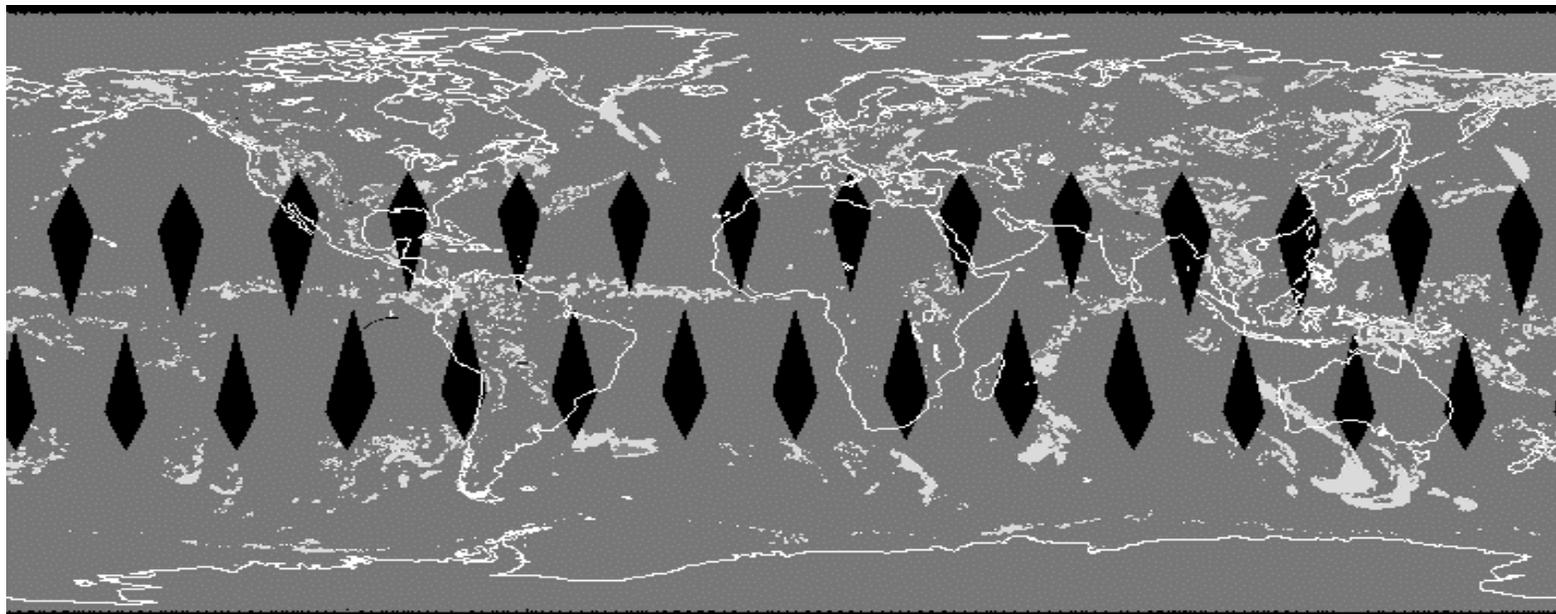


Image shows results from mining SSM/I data



Example of Data Being Mined

75 MB for one day of global data - Special Sensor Microwave/Imager (SSM/I).

Much higher resolution data exists with significantly higher volume.



Grid Miner Operations

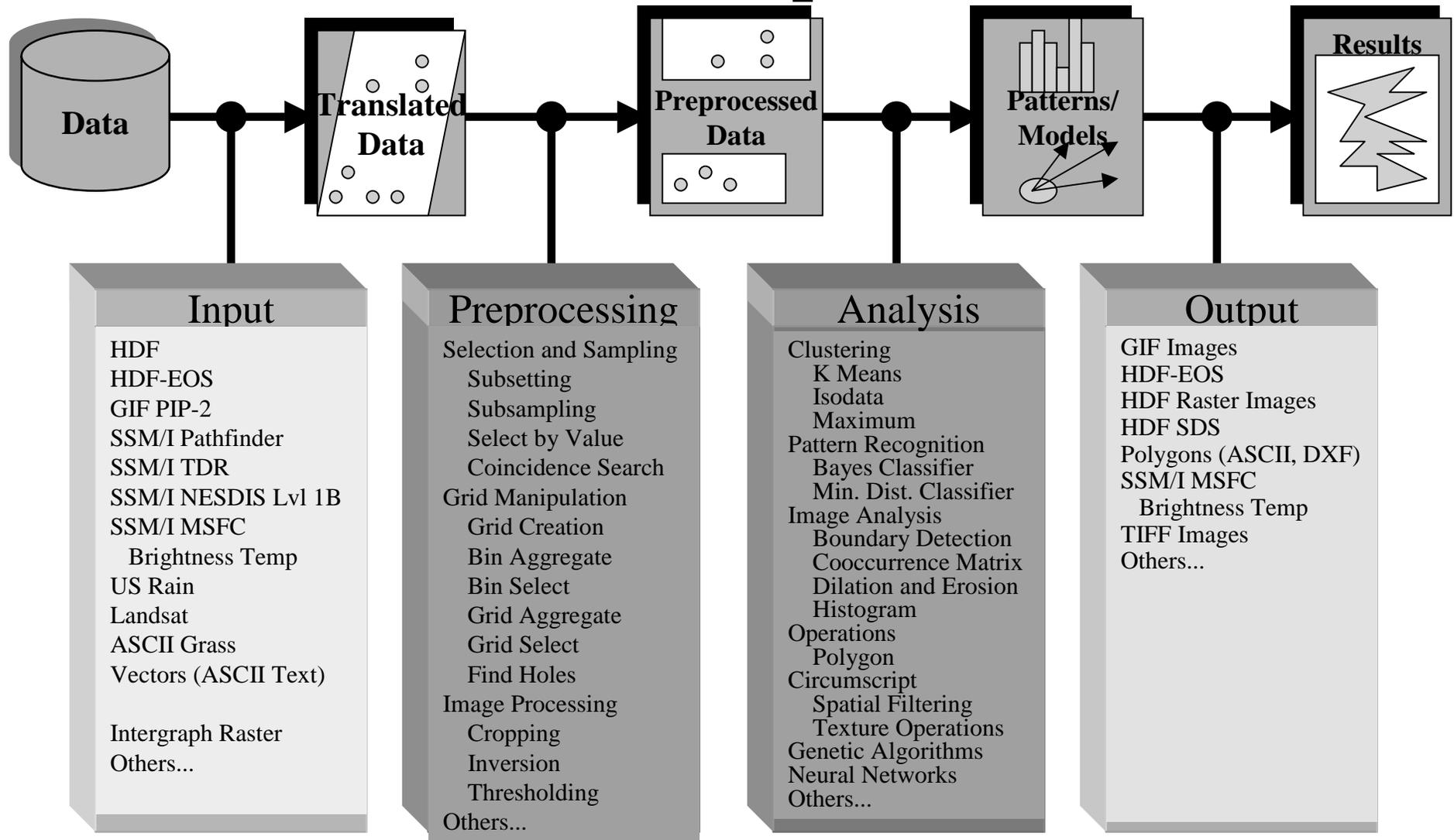


Figure thanks to Information and Technology Laboratory at the University of Alabama in Huntsville

Why Use a Grid for this Application?

NASA has large volume of data stored in its archives.

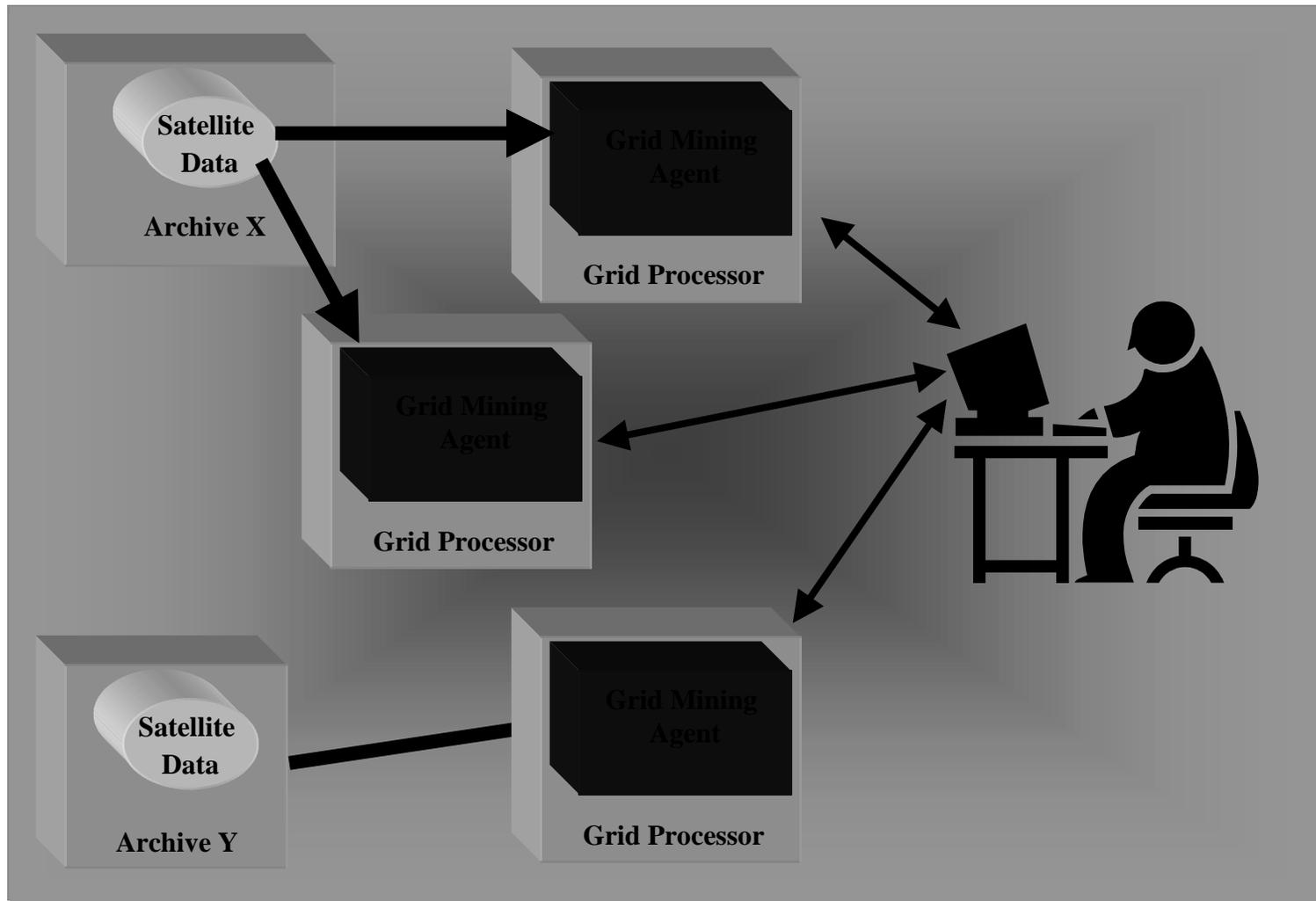
- E.g., In the Earth Science area, the Earth Observing System Data and Information System (EOSDIS) holds large volume of data at multiple archives

Data archives are not designed to support user processing

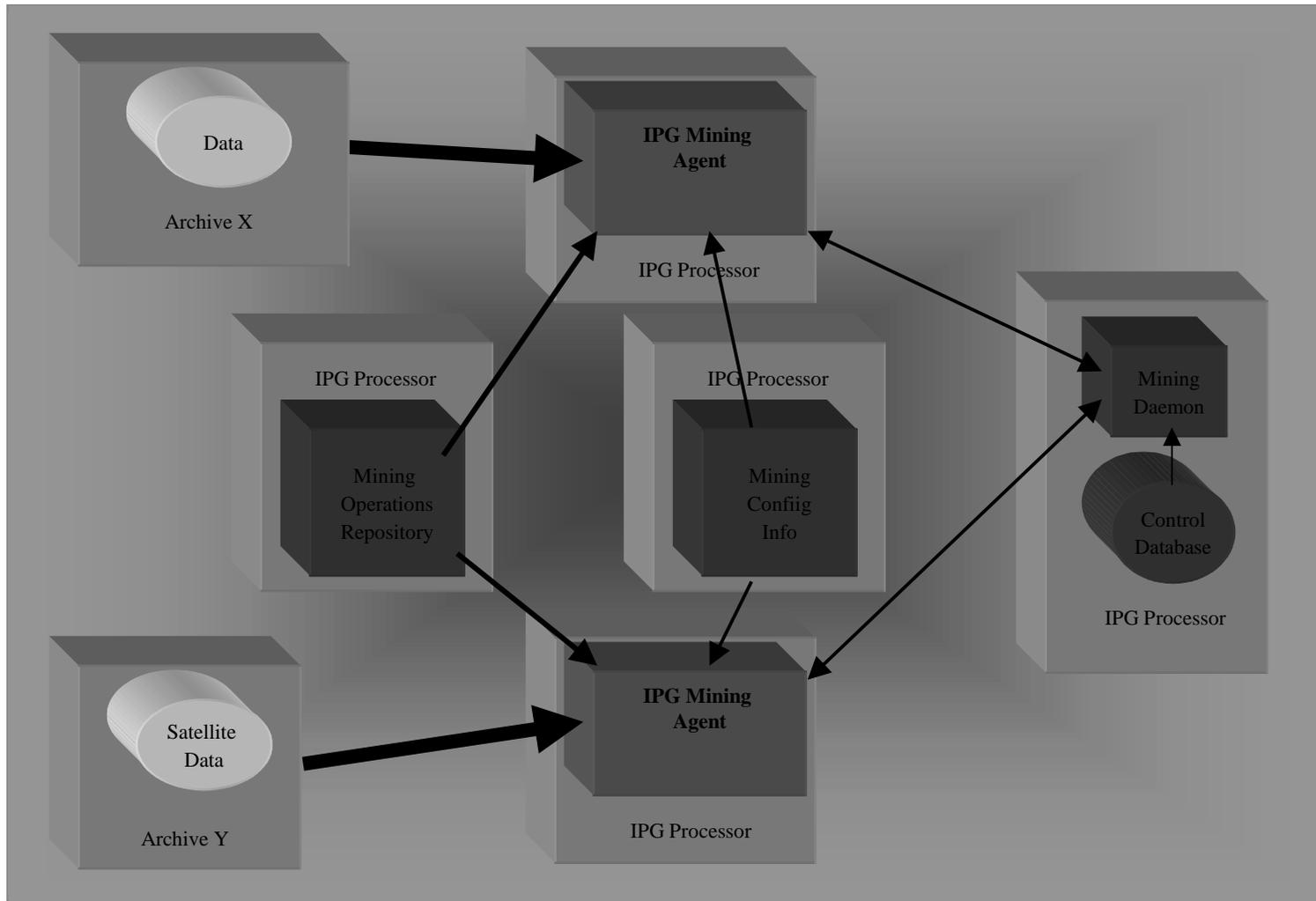
Grids, coupled to archives, could provide such a computational capability for users



Mining on the Grid



Grid Miner Architecture



Proposed Mining on the IPG

User accesses a mining portal to

- Develop mining plan
- Identify data to be mined and check file names into Control Database
- Identify nature of resources required to perform mining
- Invoke mining system

Mining portal stages N mining agents to IPG resources



Proposed Mining on the IPG

Mining agent

- Acquires configuration information from Mining Config Info server
- Acquires mining plan from mining portal
- Acquires mining operations to support mining plan using just-in-time acquisition
- Acquires URLs of data to be mined from Control Database
- Transfers data using just-in-time acquisition
- Mines data
- Sends results to specified IPG site



Mining Operator Acquisition

Vision is a number of source directories for

- Public mining operations contributed by practitioners
- For-fee mining operations from a future mining.com
- private mining operations available to a particular mining team



Starting Point for Grid Miner

Grid Miner reused code from object-oriented ADaM data mining system

- Developed under NASA grant at the University of Alabama in Huntsville
- Implemented in C++ as stand-alone, objected-oriented mining system
 - Runs on NT, IRIX, Linux
- Has been used to support research personnel at the Global Hydrology and Climate Center and a few other sites.

Object-oriented nature of ADaM provided excellent base for enhancements to transform ADaM into Grid Miner



Transforming Stand-Alone Data Miner into Grid Miner

Original stand-alone miner had 459 C++ classes.

Had to make small modifications to 5 classes and added 3 new classes

Grid commands added for

- Staging miner agent to remote sites
- Moving data to mining processor



Staging Data Mining Agent to Remote Processor

```
globusrun -w -r target_processor  
'&(executable=$(GLOBUSRUN_GASS_U  
RL)# path_to_agent)(arguments=arg1 arg2  
... argN)(minMemory=500)'
```



Moving data to be mined

```
gsincftpget remote_processor local_directory  
remote_file
```



What Grids Can Do to Support the Earth Science Community?

Can couple processing to data and data to processing

Can bring data and processing to users

Can support services of value to significant portions of the Earth Science Community

- Mining service
- Subsetting service
- Data transformation service -- from one storage format to another



What Needs to Happen for this to Become a Reality.

Data archives need to be grid-enabled

- Connected to the grid
- Provide controlled access to data on tertiary storage
 - E.g., by using a system such as the Storage Resource Broker that was developed at the San Diego Super Computer Center

Some earlier-adopter scientists need to be found to begin using the grid

Grid-enabled tools need to be made available

Sites could pool computational and data resources and form Earth Science Grid.



SRB is Existing Tool for Grid-Enabled Archive

San Diego Super Computer Center's Storage Resource
Broker (SRB).

Permits grid-access to data on tertiary storage

Supports GSI (Grid Security Infrastructure)

Provides Unix-like commands for manipulating and accessing
data

– Grid Miner uses

- `Sget -A "RESOURCE='srbresource'" pathwithfile destdir`

Datasets have logical names that are independent of location



More SRB

SRB will support data replication of a logical dataset located at different physical locations

Uses Meta data Catalog (MCAT) for holding data about the data stored in the SRB

Supports following storage systems:

- UNIX file system
- Archival storage systems such as
 - UNITREE
 - HPSS
- Large objects managed by various DBMS including
 - DB2
 - Oracle



Grid Funding

NASA is putting approximately \$7 million per year

DOE's Office of Science is putting at least \$7M/yr into Grid software development, deployment of the DOE Science Grid, and several major Grid application integration projects (high energy physics, earth sciences, fusion energy)

NSF is putting \$10-20M/yr into Grid software development and several major Grid application integration projects – e.g.

- National Earthquake Engineering Systems Grid (bring all major US earthquake engineering instruments onto a Grid)
- National Virtual Observatory (a Grid application to provide uniform access to all major astronomy datasets)

NSF is putting \$50M/yr into its new Grid based supercomputer centers (Distributed Terascale Facility)

UK eScience Grid is building a UK-wide science Grid (\$50M/yr)

European Union Data Grid (high energy physics) \$7M/yr,
EU GridLab (numerical relativity) \$3M/yr, + others



Global Grid Forum

Where did it come from

What is it

Why is it important to this community



Global Grid Forum History

Grew out of series of workshops and meetings

- Five Grid Forum workshops held between June 1999 and October 2000 in North America
 - First Workshop held at NASA Ames Research Center
- European Grid Forum (eGrid)
 - Two European Grid (eGrid) Workshops held, April 2000 and August 2000
- SC'98 and SC'99 Birds of a Feather meetings
- Middleware workshop held at Northwestern University in December 1998 with participation by Grid and Internet experts
- Grids'98: Designing, Building, and Using a National-Scale Grid", held in Chicago, July 27-28, 1998, brought together for the first time representatives of the various national Grid efforts.



Global Grid Forum Now

Represents merger of grid technical communities in North America, Europe and Asia Pacific

Meets three times per year, alternating between North America and Europe and soon Asia/Pacific

Modeled after IETF (Internet Engineering Task Force, which sets Internet standards.

Now 450 people from 35 countries working on Grid technology and standards

GGF5 meets from 21-25 July 2002 in Edinburgh, Scotland, UK



Global Grid Forum

Supports mechanism for formal review, approval and release of

- Best practices guides
- Grid standards

Organized into two types of groups

- Research Groups which coordinate research on future grid needs
- Working Groups that are expected to produce best practices documents and standards



GGF Working Groups

Grid Object Specification
(GOS)

Grid Notification Framework
(GNF)

Metacomputing Directory
Services (MDS)

Grid Security Infrastructure
(GSI)

Grid Certificate Policy
(GCP)

Advanced Reservation

Scheduling and Resource
Management

Scheduling Dictionary

Scheduler Attributes

Grid Monitoring
Architecture

Network Monitoring

JINI

NPI

OGSI

GridFTP



GGF Research Groups

Relational Database
Information Services (RDIS)
Grid Protocol Architecture
(GPA)
Accounting Models (ACCT)
Data Replication
Persistent Archives

Applications & Test beds
(APPS)
Grid User Services (GUS)
Grid Computing
Environments (GCE)
Advanced Programming
Models (APM)
Advanced Collaborative
Environments



Application & Test Beds Research Group

“The GGF Applications Research Group seeks to provide a bridge between the wider application community and the developers and directors of grid policies, standards and infrastructures.”
[APPS Web Site]

This would be one place where the Earth Science Community could inject Earth Science unique requirements into the evolving Grid development efforts.



Why is the Global Grid Forum Important to the Earth Science Community

It will result in grid standards

- It will encourage commercial products since there will be standards which the products can meet
- Products that meet accepted standards should be more marketable

It provides a forum to get Earth Science-specific requirements interjected into the grid development efforts

