Moffett Field, California, USA
NASA Ames Research Center
Thomas H. Hinkle

Application
Retrieving Earth Science Data Mining
Grids for Dummies
Outline

- Current Work
- Organization
- Background

Global Grid Forum

such as data mining

Grid support for Earth Science applications

Grids from a user's perspective

What are grids

Use of grids for applications
What are Grids?

[ widespread locat ions: ]

that are managed by diverse organizations in

disciplines, computational and information resources

software applications to integrate instruments,

These persistent environments that enable
administrative boundaries

The ability for the grid components to cross

resources

An infrastructure that is scalable to a large number

of resources

A seamless processing environment

Single sign-on

flows from resource to resource

Protection for at least authentication information as it

– public key technology

Grid Security Infrastructure (GSI), which is based

on public key technology

An underlying security infrastructure such as the

Characteristics Usually Found in Grids
Why are Grids Important?

Grids can facilitate the development of large-scale science.

• Grids provide a middleware environment that eases the development of complex systems.

• Grids provide a middle-ware environment that eases the scale science and engineering.

• Grids facilitate collaboration by providing the glue of large-scale science, instrument, and human resources through a common security model to facilitate the interaction of many different people from many different institutions.

• A common way to access and use shared data and simulations.

• A common way to access and use shared data and simulations.

• Grids provide a common way of managing distributed computing applications and systems for 21st century science, engineering, and high-performance computing and data. Grids are emerging as the infrastructure for 21st century science, engineering, and high-performance computing and data.

• Grids are processing and/or data intensive.

• Grids are widely distributed.

• Grids facilitate the development of large-scale science.
How the User Sees a Grid
capabilities with each application so that application developers do not have to re-implement this core capability. Grids will supply the core capabilities common to most applications, which will make them more efficient.

- At least some of the resources needed to solve one's problem
- Very few applications use a single computer

Support for building systems

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

Uniform data access

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

Support for collaboration

What will Grids Provide?
Uses Perl

Developed at the San Diego Super Computer Center

GridPort

Built using the Grid Portal Development Kit

Uses Java Server Pages and Java Beans

Grid Project

Developed as part of the NASA Information Power

LaunchPad

Some web portals exist for accessing grids

Web Access to the Grid is Available
Object Access Protocol

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

A set of grid functions packaged as web

- A set of grid functions

How an Application Developer Sees a Grid
Applications or faster availability of applications
Ease of development should result in more

eased with Grid functions or Grid-based web services

Will see an application whose development was

Will not see the Grid

As a user of a Grid-enabled application

All with single sign on security

Move data between Grids

Execute jobs at one or more remote sites

— Can easily

As a direct user

What a User Gains By Using a Grid
not logic of distributed interaction
Developer can focus on logic of application and
management of tasks and data
Grid functions/services handle distributed
functions
Applications can also be built by using Grid
enabled Web services.
Using capabilities provided by existing Grid
Application web services can be built by re-
What Application Developers Gain by
Specific application areas:
- Use services as building blocks to more easily develop more complex services targeting
  - Build grid services that can be used directly or
    - They will build the frameworks that allow application developers to

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

Application developers
- They will use the grid directly to realize their models and simulations.

Model builders and computational scientists
- They combine knowledge of the real world with theoretical models of the real world to
  - They will use the grid directly to realize their models and simulations.

Scientists and domain problem solvers and other users
- They need to be able to express a problem or experiment in application domain-specific
  - They will use the applications and services that the grid facilitates.

Grids support various communities or users
Summary of What User Gains

- User can focus on solving domain issues of the problem and not on computer science issues of distributed computing
A number of universities - San Diego Supercomputer Center
National Center for Supercomputing Applications (NCSA)
Advanced Computational Infrastructure Programs at
National Science Foundation PACI (Partnerships for
NASAs Ames Information Power Grid Team
Science Institute
University of Southern California Information
Argonne National Laboratory
development personnel from

The Globus project involves research and
such example

Most Grids Are Built on the Globus Toolkit

NASAs Information Power Grid (IPG) is one
Data Mining on the Grid

- A proposed IPG Mining Service
- Information Power Grid (IPG)
- The Grid Miner developed for NASA's
- Why mine on the Grid?
- What is data mining?
to Scientific Data, Oct 1999, on Issues in the Application of Data Mining search through the data. "NASA Workshop using techniques that go beyond a simple "What is Data Mining?"

Data mining is the process by which...
Grid Miners

- Provides basis for what could be an on-going Grid Mining Service
- Provided basis for satisfying one of two major IPG milestones last year
- Helped debug the IPG on the IPG
- Developed as one of the early applications
Image shows results from mining SSM/I data

Example: Mining for Mesoscale Convective Systems
significantly higher volume.

Much higher resolution data exists with

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

Example of Data Being Mined
computational capability for users.

* Grids, coupled to archives, could provide such a processing

**Data archives are not designed to support user

Large volume of data at multiple archives

System Data and Information System (EDSIS) holds

E.O. In the Earth Science area, the Earth Observing

archives.

NASA has large volume of data stored in its

Why Use a Grid for this Application?
Mining on the Grid
Grid Miner Architecture
Proposed mining on the IPC

Mining portal stages N mining agents to
invoke mining system
perform mining

Identity nature of resources required to

Identity data to be mined and check file names

Develop mining plan

User accesses a mining portal to

IPG resources
Proposed mining on the IPC

- Sends results to specified IPC site
- Mines data
- Transfers data using just-in-time acquisition
- Acquires URLs of data to be mined from Control Database
- Time acquisition
- Acquires mining operations to support mining plan using just-in-
  time acquisition
- Acquires mining plan from mining portal server
- Acquires configuration information from Mining Config Info
- Mining agent
particular mining team
private mining operations available to a
mining.com
for-fee mining operations from a future
practitioners
Public mining operations contributed by
Vision is a number of source directories for

Mining operator acquisition
For enhancements to transform ADAM into Grid Miner

- Object-oriented nature of ADAM provided excellent base for enhancements
- Hydrology and Climate Center and a few other sites
- Has been used to support research personnel at the Global Run on NT, IRIX, Linux
- Runs on NT, IRIX, Linux
- System
- Implemented in C++ as stand-alone, object-oriented mining
- Huntsville
- Developed under NASA grant at the University of Alabama in mining system
- Grid Miner reused code from object-oriented ADAM data

Starting Point for Grid Miner
Moving data to mining processor
Staging miner agent to remote sites

Grid commands added for
classes and added 3 new classes

Had to make small modifications to 5
classes.

Original stand-alone miner had 459 C++
...\( \text{argN} \) (minMemory=500)
\( \text{RL} \) # path-to-agent (arguments=arg1 arg2
executable=$GLOBUSRUN CLASS
globusrun -w -r target-processor

Staging Data Mining Agent to Remote Processor
Moving data to be mined

- gsincftpget remote_processor
  local_directory remote_file
Another Data Transformation Service -- from one storage format

- Subsetting Service
- Mining Service

 Portions of the Earth Science Community

Can support services of value to significant

- Can bring data and processing to users
- Processing
- Can couple processing to data and data to

Earth Science Community?

What Grids Can Do to Support the
and form Earth Science Grid.
Sites could poor computational and data resources
- Grid-enabled tools need to be made available
- begin using the grid
- Some earlier-adopter scientists need to be found to
- that was developed at the San Diego Super Computer Center
  - E.g., by using a system such as the Storage Resource Broker
  - Provide controlled access to data on tertiary storage
  - Connected to the grid
  - Data archives need to be grid-enabled

Become a Reality.

What Needs to Happen for this to
Datasets have logical names that are independent

Grid Miners use

- Grid Miners
- and accessing data

Provides Unix-like commands for manipulating

Supports GSI (Grid Security Infrastructure)

Permits grid-access to data on tertiary storage

Resource Broker (SRB)

San Diego Super Computer Center's Storage

SRB is Existing Tool for Grid-Enabled Archive
Oracle •
DB2 •
Large objects managed by various DBMS including
HPS •
UNITE •
Archival storage systems such as
UNIX file system •
Supports following storage systems:
about the data stored in the SRB
Use Meta data Catalog (MCAT) for holding data
dataset located at different physical locations
SRB will support data replication of a logical
More SRB
EU GridLab (numerical relativity) $3M/yr + others

European Union Data Grid (high energy physics) $7M/yr

UK Science Grid is building a UK-wide science Grid ($250M/yr)

(Distributed Terascale Facility)

NSF is putting $50M/yr into its new Grid-based supercomputer centers

major astronomy datasets

National Virtual Observatory (a Grid application to provide uniform access to all
current and future observatories on a Grid)

Earthquake Engineering Research Centers Grid (bring all major US earthquake

Grid application integration projects - e.g.

NSF is putting $10-20M/yr into Grid software development and several major

energy

application integration projects (high energy physics, earth sciences, fusion
development, deployment of the DOE Science Grid, and several major Grid

DOE's Office of Science is putting at least $7M/yr into Grid software

NASA is putting approximately $7 million per year

Grid Funding
Global Grid Forum

Why is it important to this community •
What is it •
Where did it come from •
national Grid efforts.

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

Global Grid Forum History
Scotland, UK

- GGF5 meets from 21-25 July 2002 in Edinburgh.

- Grid technology and standards
  Now 450 people from 35 countries working on
  Force, which sets Internet standards.

- Modeled after IETF (Internet Engineering Task
  North America and Europe and soon Asia/Pacific
  Meets three times per year, alternating between
  North America, Europe and Asia-Pacific

- Represents merger of grid technical communities

Global Grid Forum Now
practices documents and standards
Working Groups that are expected to produce best
Grid needs
Research Groups which coordinate research on future
Organized into two types of groups
- Grid standards
- Best practices guides
and released of
Supports mechanism for formal review, approval

Global Grid Forum
CGF Working Groups

- CGF
- OGS
- NPI
- INI
- Network Monitoring
- Architecture
- Grid Monitoring
- Scheduler Attributes
- Scheduling Dictionary
- Management
- Scheduling and Resource

- Advanced Reservation
- (GCP)
- Grid Certificate Policy
- Infrastructure (GSI)
- Grid Security
- Services (MDS)
- Metacomputing Directory
- Framework (GNE)
- Grid Notification
- (GOS)
- Grid Object Specification

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

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

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

[AQPS Web Site]

Grid policies, standards and infrastructures.
community and the developers and directors of
provide a bridge between the wider application
The CGF Applications Research Group seeks to

Group

Application & Test Beds Research
grid development efforts into the specific requirements intercepted by the
- It provides a forum to get Earth Science
  - Products that meet accepted standards should be more marketable
  - There will be standards which the products can meet
  - It will encourage commercial products since it will result in grid standards

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