Featuring Earth Science Data Mining

Grids for Dummies

Thomas H. Hinke

NASA Ames Research Center
Moffett Field, California, USA
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
[widespread locations: https://www.globus.org]

that are managed by diverse organizations in disciplines, computational and information resources
software applications to integrate instruments,
"Grids are persistent environments that enable"

What are Grids?
Middleware Makes the Grid
administered boundaries
The ability for the grid components to cross
resources
An infrastructure that is scalable to a large number
A seamless processing environment
Single sign-on
flow from resource to resource
Protection for at least authentication information as it
on public key technology
Grid Security Infrastructure (GSI), which is based
An underlying security infrastructure such as the

Characteristics Usually Found in Grids
Why are Grids Important?

- Grids provide a middleware environment that eases the scale science and engineering.
- Grids facilitate collaboration by providing the glue of large scale, instrument, and human resources applications and systems.
- Grids facilitate the development of large-scale science.
- Grids can facilitate the development of large-scale science.
- Grids provide a common way of managing distributed computing.
- 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.
- Grids provide a common way of identifying distributed computing.
How the User Sees a Grid

- A set of grid functions that are available as
  - Command-line functions
  - Application programmer interfaces (APIs)

- After authentication, functions can be used to
  - Spawn jobs on different processors with a single command

- Support the communication between programs executing on different processors
- Access data on remote systems
- 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).
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.

At least some of the resources needed to solve one's problem invariably reside elsewhere.

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 (SDL)
- Interface is defined through WSDL

- A set of grid functions packaged as web services
- A set of grid functions

How an Application Developer Sees a Grid
Ease of development should result in more

As a user of a grid enabled application

- 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.

Applications web services can be built by re-
using capabilities provided by existing grid.

Using Grids
What Application Developers Gain by
specific application areas:

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

**Service builders**

- They will use models and simulations as components.
- Compiling or a large number of distributed processes.
- 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.
- They will use the grid directly to realize their models and simulations.

**Scientists and domain problem solvers and other users**

- They will use the applications and services that the grid facilitates.
Issues of distributed computing
the problem and not on computer science
User can focus on solving domain issues of
Summary of What User Gains
A number of universities
- San Diego Supercomputer Center
- National Center for Supercomputing Applications (NCSA)
- Advanced Computational Infrastructure (Partnerships for
  National Science Foundation PACI (Partnerships for
  - NASA's 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
  NASA's Information Power Grid (IPG) is one
Most Grids Are Built on the Globus Toolkit
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?

Data Mining on the Grid
What is Data Mining


To Scientific Data, Oct 1999,  

an Issues in the Application of Data Mining  

search though the data. "NASA Workshop  

using techniques that go beyond a simple  

from a potentially large volume of data  

information and knowledge are extracted  

Data mining is the process by which  

•
Grid Miner

Going Grid Mining Service
- Provides basis for what could be an on-
  IPG milestones last year — Provided basis for satisfying one of two major
  - Helped debug the IPG on the IPG
  Developed as one of the early applications

Grid Miner
Image shows results from mining SSM/I data

Example: Mining for Mesoscale Convection Systems
Much higher resolution data exists with the Sensor Microwave Imager (SMI). A 75 MB file contains areal data for one day of global data.
Grid Miner Operations
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 (EOSDIS) holds

-E.g., 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
Proposed mining on the IPC

• IPC resources

Mining portal stages

• Invoke mining system
• Perform mining
• Identify nature of resources required to
  control database
• Identify data to be mined and check file names
• Develop mining plan

User access to mining portal

•
Proposed mining on the IPC

- Sends results to specified IPC site
- Mines data
- Transfers data using just-in-time acquisition
- Acquires URL's or data to be mined from Control Database
- Time acquisition
- Acquires mining operations to support mining plan using just-in-
  server
- Acquires mining plan from mining portal
- Acquires configuration information from Mining Configuration
  Info
- Mining agent
Minning operator acquisition

Vision is a number of source directories for

Particular mining team - private mining operations available to a
mining.com

For-fee mining operations from a future
practitioners

Public mining operations contributed by
Grid Miner

Starting Point for Grid Miner

- Object-oriented nature of ADAM provided excellent base
- Hydrology and Climate Center and a few other sites
- Has been used to support research personnel at the Global Run 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
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 classes.
- Original standalone miner had 459 C++

Miner Into Grid
Transforming Stand-Alone Data
Staging Data Mining Agent to Remote Processor
Moving data to be mined

Local directory
Remote File

Remote processor

Gnashpipe
to 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
- 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-adapter 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

Set a "RESOURCE="string"" path with the destdir

Grid Miner uses

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

HPS2
UNETREE

Archival storage systems such as

UNIX file system

Supports following storage systems:

about the data stored in the SRB

Uses Meta data Catalog (MCAT) for holding data

dataset located at different physical locations

SRB will support data replication on 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 ($50M/yr)

(Distributed Terascale Facility)

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

major astronomy datasets

National Virtual Observatory (a Grid application to provide uniform access to all

existing instruments in a Grid)

Earthquake Engineering Systems 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 meetings held in Chicago, July 27-28, 1998, brought
- Grids'98: Designing, Building, and Using a National-
  Internet experts
in December 1998 with participation by Grid and
- Middleware Workshop held at Northwestern University
  SC'98 and SC'99 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

GF2 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.

Modelled 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 release of
  Supports mechanism for formal review, approval

Global Grid Forum
GCF Working Groups

- GridFTP
- OGSI
- NPI
- INI
- Network Monitoring
- Architecture
- Grid Monitoring
- Scheduler Attributes
- Dictionary
- Management
- Scheduling and Resource
- Advanced Reservation (GCP)
- Grid Certificate Policy Infrastructure (GSI)
- Grid Security Services (MDS)
- Metacomputing Directory
- Framework (GND)
- Grid Notification (GOS)
- Grid Object Specification
Research Groups

Environments
Advanced Collaborative Models (ACM)
Advanced Programming Environments (ACE)
Grid Computing
Grid User Services (GUS)
Applications & Test beds

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

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

[APPS Web Site]

Grid policies, standards and infrastructures.

community and the developers and directors of
provide a bridge between the wider application
The CGF Application Research Group seeks to

Group

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

Why is the Global Grid Forum Important Community