Moffett Field, California, USA

NASA Ames Research Center

Thomas H. Hinkle

Application

Retrieving Earth Science Data Mining

Grids for Dummies
Current Work
Organization
Background

Global Grid Forum

Such an as data Mining
Grid Support for Earth Science applications

Grids from a user’s perspective

What are grids

Use of grids for applications

Outline
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/]
administrative boundaries

The ability for the grid components to cross

resources

An infrastructure that is scalable to a large number

processing environment

Single sign-on

Flows from resource to resource

Protection for at least authentication information as it

on public key technology

Grid Security Infrastructure (GSI), which is based

underlying security infrastructure such as the

Characteristics Usually Found in Grids
Why are Grids Important?

- Grids provide a middleware environment that eases the development of complex systems.
- Grids provide 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.
- A common way to access and use shared data and simulations.
- Grids facilitate collaboration by providing the glue of large-scale science.
- Grids provide a common way of managing distributed computing applications and systems.
- Grids enable 21st century science, engineering, and high-performance computing and data. Grids are emerging as the infrastructure for processing and/or data-intensive applications.
- Grids can facilitate the development of large-scale science.
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
  - Move data from one processor to another
  - 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).
 capability with each application so that application developers do not have to re-implement this core.

Grids will supply the core capabilities common to most applications, inevitably reside elsewhere.
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?
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

Uses Perl
Object Access Protocol (OAP) is SOAP (Simple Object Access Protocol) — Standard access protocol is SOAP (Simple Object Access Protocol) — Services Description Language (SDL) — Interface is defined through WSDL (Web Services Description Language) • A set of grid functions packaged as web services • 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 sites

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

- Grid functions/services can also be built by using Grid
- Applications can be built by reusing capabilities provided by existing Grid
- Application web services can be built by reusing Grid

Using Grids

What Application Developers Gain by
Specific application areas:

- Use services as building blocks to more easily develop more complex services leveraging
  service-oriented architecture.
- Build grid services that can be used directly or as components.
- They will build the frameworks that allow application developers to
  Service builders
  - They will use the models and simulations as components.
  - Compiling or debugging 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.
- They will produce simulations of models that can produce a complete representation of the
  world, theoretical models of the real world to
  - Model builders and computational scientists
  - They will use the grid directly to realize their models and simulations.
  - They need to be able to express a problem or experiment in application domain-specific
    terms, specifically the drivers (initial conditions, input data sources, etc.) request that the solution
    be obtained, and manage the resulting graphics, tables, etc.
  - They will use the applications and services that the grid facilitates.

Grids Support Various Communities of Use
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 Programs at
  National Science Foundation PAG (Partnerships for
  NASA's Ames Information Power Grid Team
- Ames Institute
- University of Southern California's Information
  Development Personnel from
  The Gloubs project involves research and
  such example
  NASA's Information Power Grid (IPG) is one

Most Grids are built on the Gloubs 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
Grid Miner Service
- Provides basis for what could be an on-
- IPG milestones last year
- Developed as one of the early applications
Example: Mining for Mesoscale Convective Systems
• 75 MB for one day of global data - Special Sensor Microwave/Imager (SSMI).

• Much higher resolution data exists with significantly higher volume.

Example of Data Being Mined
Figure thanks to Information and Technology Laboratory at the University of Alabama in Huntsville.
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.O.s, 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
IPG resources

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

Proposed mining on the IPG
Proposed mining on the IPC

- Sends results to specified IPC site
- Mines data
- Transfers data using just-in-time acquisition
- Acquires URL's of data to be mined from Control Database
- Time acquisition
- Acquires mining operations to support mining plan using just-in
- Acquires mining plan from mining portal
- Server
- Acquires configuration information from Mining Control 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 directions for

Mining operator acquisition
Grid Miner

Starting Point for Grid Miner

- Object-oriented nature of ADAM provided excellent base
- Hydromology and Climate Center and a few other sites
- Has been used to support research personnel at the Global System
  - Runs on NT, IRIX, Linux
  - 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 5 classes.
Original stand-alone miner had 459 C++

Miner into Grid Miner
Transforming Stand-Alone Data
globusrun -w -r target_processor
& (executable=$(GLOBUSRUN_GASS_U
  RL) # path_to_agent)(arguments=arg1 arg2 ...
argN)(minMemory=500)

Staging Data Mining Agent to Remote Processor
Local directory remote file
get remote processor

Moving data to be mined
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-adaptor 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

– Provides 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="superresource" pathwhitelist descriptor.

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:

HPS

UNITEE

Archival storage systems such as:

UNIX file system

Supports following storage systems:

about the data stored in the SRB

Uses Meta data Catalog (MCAST) 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 ($50M/yr)

(Distributed Terascale Facility)

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

major astronomy datasets

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

engineering instruments onto a Grid)

National Earthquake Engineering Systems Grid (bring all major US earthquake

Grid application interaction projects)

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

(energy)

application interaction 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: 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
- First workshop held at NASA Ames Research Center
- and October 2000 in North America
- Global 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 ~250 people from 35 countries working on
- Internet Engineering Task Force, which sets Internet standards.

- Modeled after IETF
- 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**
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
GF Working Groups

- GridFTP
- OSGI
- 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 (GND)
- Grid Notification Specification (GOS)
- Grid Object Specification
Environments
Advanced Collaborative Models (APC)
Advanced Programming Environments (GCE)
Grid Computing
Grid User Services (GUS)
Applications & Test beds

GCF Research Groups

- 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 intersect Earth Science unique
This would be one place where the Earth Science

[APPS Web Site]

GRID policies, standards and infrastructures.

and the developers and directors of
provide a bridge between the wider application

The GGF Applications Research Group seeks to

Group

Application & Test Beds Research
grid development efforts
specific requirements intersected 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 to the Earth Science Community