Moffett Field, California, USA

NASA Ames Research Center

Thomas H. Hinke

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?

[wide spread locations: [http://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..."

*
Middleware Makes the Grid
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

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 facilitate collaboration by providing the glue of large-scale science, engineering, and operational applications
- grids can facilitate the development of large-scale science and engineering
- grids provide a middleware environment that eases the development of complex systems
- 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 managing distributed computing
- grids provide a common way of managing distributed resources
- data, instrument, and human resources
- for 21st century science, engineering and high-performance computing and data, grids are emerging as the infrastructure

... (additional text)

- grids are widely distributed
- grids are processing and/or data intensive
soon

then negotiate the reservation and execution (coming)

– Use a broker to select the best place for a job to run and

available on the Grid using the grid information service

– Discover the properties of computational resources

executing on different processors

– Support the communication between programs

move data from one processor to another

– Access data on remote systems

command

– Spawn jobs on different processors with a single

After authentication, functions can be used to

Command-line functions

Application programmer interfaces (APIS)

A set of Grid functions that are available as

How the User Sees a Grid
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 will 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 Supercomputer Center

GridProject

Built using the Grid Portal Development Kit

Uses Java Servlet Pages and Java Beans

Grid Project

Developed as part of the NASA Information Power

LaunchedPad

Some web portals exist for accessing grids

Web Access to the Grid is Available
Object Access Protocol (SOAP)

- Standard access protocol is SOAP (Simple Object Access Protocol)

Web Services Description Language (WSDL)

- Interface is defined through WSDL (Web Services Description Language)

A set of grid functions packaged as web services

A set of grid functions sees a Grid

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
  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-

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

Model builders and computational scientists:

- They need to be able to express a problem or experiment in application domain-specific
- 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
  - Terms, specify the drivers (initial conditions, live data sources, etc.) request that the solution

Scientists and domain problem solvers and other users:

Grids Support Various Communities of Use
Summary of What User Gains

User can focus on solving domain issues of the problem and not on computing issues of distributed computing.
A number of universities - San Diego Supercomputer Center
NCSA (National Center for Supercomputing Applications)
Advanced Computation Infrastructure Programs at
National Science Foundation PACI (Partnerships for
Science Institute)
NASA's Ames Information Power Grid Team
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

SciMin: Data, Oct 1999, Application of Data Mining

on Issues in the Application of Data Mining

search through the data. NASW 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

What is Data Mining
Grid Miner Service

- 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.
significantly higher volume.

- Much higher resolution data exists with Simplified Sensor Microwave Imager (SSMI).

- 75 MB for one day of global data - Special Example of Data Being Mined
Why Use a Grid for This Application?

• 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.g., in the Earth Science area, the Earth Observing
  Archives.

• NASA has large volume of data stored in its
Mining on the Grid
Grid Miner Architecture
IPG resources

Mining portal stages N mining agents to

Invoke mining system

Perform mining

Identity nature of resources required to

Into Control Database

Identity data to be mined and check file names

Develop mining plan

User access to 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 UI/UX of data to be mined from Control Database
- Acquires mining operations to support mining plan
- Acquires mining plan from mining portal
- Acquires configuration information from Mining Console Info
- Mining agent •
particular mining team
- private mining operations available to a
  mining.com
- For-free 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

Hydrology and Climate Center and a few other sites.

Has been used to support research personnel at the Global

- Runs on NT, IRIX, Linux

system

- Implemented in C++ as stand-alone, object-oriented mining

system

- 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
Grid 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
... args(minMemory=500),

RL # path-to-agent)(arguments=arg1 arg2
executable=$GLOBUSRUN CLASS
.globusun -w -r target-processor

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

local directory remote file

get ciphertext remote processor
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
- 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-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

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 of location.

- Set -A "RESOURCE=lsresources pathwithfile 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:
HPSS
UNITEE
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 or 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 $5OM/yr into is new Grid based supercomputer centers
- major astronomy datasets
- National Virtual Observatory (a Grid applicable to provide uniform access to all
  instruments on a Grid)
- Earthquake Engineering Systems Grid (bring all major US earthquake
  Grid application integration projects - e.g.
  NSF is putting $10-2OM/yr into Grid software development and several major
  projects)

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**
Why is it important to this community

• What is it

• Where did it come from

Global Grid Forum
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
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

GGF meets from 21-25 July 2002 in Edinburgh,

Grid technology and standards

Now 450 people from 35 countries working on

Now A 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 release of
  - Supports mechanism for formal review, approval

Global Grid Forum
GridFTP
OGSI
NPI
INII
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 (GPF)
Grid Notification (GOS)
Grid Object Specification

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

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

GF Research Groups
Application & Test Beds Research Group

"The GGF Applications Research Group seeks to provide a bridge between the wider application 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.
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  
result in grid standards  

Why is the Global Grid Forum Important

Community