Building Grid-enabled Applications in Bioinformatics and Digital Archive

Eric Yen, Horng-Chun Lee, William Ueng, Simon Lin
Computing Centre, Academia Sinica

28 July 2004
Goals of Grid Development in AS

• Take Advantage of Grid Technology to
  – Facilitate resource sharing and collaboration in Taiwan and with international academic institutes
  – Build up more robust IT Infrastructure
  – Federating distributed resources of computing, storage and data

• Learn from LCG/HEP to expand to other academic discipline, such as bioinformatics, virtual observatory, astronomy, biodiversity, digital archive, and toward eScience

• Provide Secure, Reliable and Ubiquitous Services
Building Grid Applications
Grid-enabled Application

- Not just can run in a grid, also should take advantage of the virtualized grid infrastructure to accelerate processing time or to increase remote collaboration.

- In terms of Grid Service, grid enablement means that the application can run as a Web/Grid service in a grid environment, while making use of the various services provided by the grid infrastructure.

- Applications must be accessible as Web/Grid Services.
Grid Application Development

• Provides Toolkits or Grid Services in a way that the end-users and especially application developers can build and run applications on the Grid without needing to know details about the runtime environment in advance.

• To simplify distributed heterogeneous computing in the same way that the World Wide Web simplified information sharing over the Internet.

• Grid-enable means -- different parts of the application can be run simultaneously at different locations.
Grid Application Framework

- LCG Application Area
- GrADs: Grid Application Development Software Project
- GridLab and GridSphere: A Grid Application Toolkit and Testbed
- IBM Grid Application Framework for Java (GAF4J)
LCG Application Area

- Scope includes common applications software infrastructure, frameworks, libraries, and tools; common applications such as simulation and analysis toolkits; grid interfaces to the experiments; and assisting the integration and adaptation of physics applications software in the grid environment.

- Projects
  - PI
  - POOL/CondDB
    - SEAL
      * Simulation
      * SPI
GridLab Project

- Funded by the EU (5+ M€), January 2002 – March 2005
- Application and Testbed oriented
  - Cactus Code, Triana Workflow, all the other applications
- Main goal: to develop a Grid Application Toolkit (GAT) and set of grid services and tools (GridSuite):
  - Resource management (GRMS),
  - Data management (GDMS),
  - Monitoring (Mercury) and information services,
  - Adaptive components (Pythia),
  - Mobile user support and remote visualization,
  - Security services (GAS),
  - Portals (GridSphere),
- ... and test them on a real testbed with real applications

Visit to AIST, Tokyo, Japan 23 April, 2004
GridLab Architecture

GridLab Services
- GAS
- Delphoi
- Viz Service
- GRMS
- Replica Catalog
- Data Movement
- Mobile Services
- Mercury

GridSphere Portal
- Mobile clients
- Application Layer
  - Astrophysics
  - Bioinformatics
  - Other Life Sciences
  - ...

GAT API
- GSI-enabled Web Service API
- Third Party Services and Libraries
- C and Java APIs to Globus 2.x/3.x Pre-WS and other Core Services

GridLab Testbed and Infrastructure
Portal standards

- **JSR 168 Portlet API ratified August 2003**
  - Similar to Servlet API in providing reusable web applications
  - Ratified by vendors including BEA, Sun, IBM, Oracle, Plumtree and others...

- **WSRP (Web Services for Remote Portlets) ratified by OASIS committee**
  - Specifies how web services can be consumed by standards compliant portals

- **Java Server Faces ratified**
  - Specifies an event based user interface for web presentation development
The goal is to develop the program development and execution environment required to make performance on the Grid truly accessible for scientists and engineers.

Project has proceeded using phased research and development strategy.

- Integrating mature and evolving software
- Addressing 1-10 year research problems
- Focusing on software development for the most complex, dynamic and heterogeneous computational platform to date

http://hipersoft.cs.rice.edu/grads/
The Basic GrADS Software Architecture

Program Preparation System

- Source application
- whole program compiler
- Config. object program
- Software components
- libraries

Performance feedback

Execution Environment

- Scheduler/Service Negotiator
- Realtime perf monitor
- Dynamic optimizer
- Grid runtime System (Globus)

- Perf problem
- negotiation
IBM Grid Application Framework for Java (GAF4J)

- A lightweight framework that abstracts all grid semantics from the application logic and provides a simpler programming model that lines up smoothly with common Java™ programming models.
Strategy for Grid Enablement

David Kra
Strategy for Grid Application Enablement
IBM developWorks
Types of Grid AP Enablement (1)

Strategy 1: Batch Anywhere

- Application Instance
- Grid Node

Strategy 2: Independent Concurrent Batch

- Application Instance 1
- Application Instance 2
- Application Instance 3

David Kra
Strategy for Grid Application Enablement
IBM developWorks
Strategy 3: Parallel Batch
Parallel Batch takes each user's batch work, subdivides it, disperses it out to multiple nodes, collects it, and then aggregates the results.

Strategy 4: Service
transition from a batch to a service-oriented architecture

David Kra
Strategy for Grid Application Enablement
IBM developWorks
Types of Grid AP Enablement (3)

Strategy 5: Parallel Services
combines the service-oriented architecture of Strategy
4: Service with the subdivided work model of Strategy
3: Parallel Batch

Strategy 6: Tightly Coupled Parallel Program
provides intense communications and synchronization:
• Between client and services
• Among services

David Kra
Strategy for Grid Application Enablement
IBM developWorks
Grid Service Management

- Grid services are (extended) Web services:
  - Can use Web service management interfaces
  - Additional interfaces for Grid services being defined

- Grid service capabilities:
  - Service lifecycles
  - State values

- Special infrastructure services must be managed:
  - Handle Resolvers
  - Factories
  - Registries
  - Program Execution services
  - etc.
Grid Technologies

- Grid Portals - GridPort
- Workflow control pipelines - Chimera/Pegasus
- Job scheduling management - CondorG
- Job execution system - GRAM
- Data caching and replication - RLS
- Authentication system - GSI
- Large file data transport – GridFTP, RFT
- Metadata catalog - MCS, MCAT
- Collection management – SRB
- Database Access on the Grid: OGSA-DAI
Challenges

• Grid technology is rapidly evolving; activities in progress
  – WSRF-based Grid Service
  – GridFTP rewrite, protocol redesign
  – Virtual Data System redesign (support collection-based access)
  – OGSA-DAI data access interface
  – Data Format Description Language
  – Replica Management
  – Grid File System
  – Interoperability
  – ...

Bioinformatics Grid
Challenge of Bioinformatics (1)

• Integration of many different sources of data
  – Not consistent metadata

• Validation or correction of experimental data

• To hide the complexity and provide transparent access to the Grid services

• However, Grid is still largely a framework, explicit support to Bioinformatics and Contents needs to be worked out
Challenge of Bioinformatics (2)

• Life Science is a data-driven science
  – Data is the key issue of bioinformatics

• Most of LS applications (i.e. workflows) are built based on try-and-error processes
  – It may change rapidly because of researchers’ purpose
  – Dynamic workflow is required

• Most of LS researchers prefer an intuitive graphic user interface instead of command line options
  – Web based portal is required

• Heterogeneous computing resources need to be integrated and shared coordinately
  – Grid is a dynamic system for resource sharing
Bioinformatics Grid Service Infrastructure

Bioportal Middleware

Web/Grid Services

BioGrid Services

OGSA/OGSI

Web/Grid Services

Computing Resources

Target System (Super Computer)

Target System (PC-Cluster)

Data Resources

Target System (Database System)
Core BioGrid Services

• Workflow Management Service
  – Workflow handling

• User Management Service
  – User Authentication and Single-Sign-On

• Resource Management Service
  – User Authorization
  – Application level Resource Broker
  – Application information collector and manager

• Job Management Service
  – Abstraction layer of computing elements

• Data Management Service
  – Abstraction layer of storage elements
The Portal

• **Should be ...**
  – easy to use and be available everywhere
    • Intuitive web interface
  – able to deploy user defined workflow
    • workflow (application) container

• **Core technology**
  – Java and XML
System Architecture - The hierarchy

- 3-Layer hierarchy
- Grid-enabled environment
- Temporary solution for grid middleware
  - Virtual Queue (VQ)
  - Computing Element Metadata Manager (CEMM)
  - Local System Agent (LSA)
Design Considerations

- Allows for the end user to share the computing resources and complex databases under one system
- The scale of computing resource can be re-sized as needed
- The computing resource can be allocated dynamically
- The complexities of computing resources allocation should be hidden by web interface
Grid Enabled BioPortal@ASCC

• Implementation
  – A Web-based uniform entrance for providing bioinformatics computing service to biology researchers world-wide
  – The integration of heterogeneous computing platforms
  – The integration of federated bioinformatics data bases
  – A high throughput computing environment

• Current Status
  – Open for On-line services since July 2003
  – Based on home grown grid emulated middleware
  – On-line applications including
    ➢ Analysis Tools: NCBI BLAST (all types), CRASA, FASTA, OPASS, R (Microarray analysis), etc.
    ➢ Databases: SMD
  – Web Site: http://bits.sinica.edu.tw
CRASA: Complexity Reduction Algorithm for Sequence Analysis

Complexity Reduction Algorithm for Sequence Analysis (CRASA) is a homology-based utility tool for the annotation and analysis of large genomic sequences. This tool is a collaboration product of Computing Centre and Biomedical Institute in Academia Sinica.

The human genome has been sequenced near completion by the international sequencing consortium. From the massive data deposited in the public database, it is possible to annotate the estimated 60,000 to 100,000 human genes in the genome. Gene prediction is done conventionally by two different, but complementing, statistics- and homology-based approaches, as demonstrated in annotating the eukaryotic genomes of Caenorhabditis elgans and Drosophila melanogaster.
Digital Archive Grid
Challenges for DL/M/A

• Fragmentation of existing data resources and assets due to
  – A heterogeneous environment
  – Under-utilized computing and storage resources
• Cumbersome data access and poor integration
• Data security and protection
• Complex management of decentralized systems and resources
• High total costs of IT infrastructure. Inflexible and difficult-to-change system
• Develop Grid Services that can integrate heterogeneous metadata systems, distributed database management systems and geospatial information systems.
• Provide a framework to exchange different xml documents (EAD, DC ...) in “National Digital Archives Program”.
Workflow and Architecture for Digital Archive

Originals Of Project

Non-Digital

Digital

Digitalization

Verification

Content Analysis

Archive Creation

Post-Proc

Digital Entity(P1)

Digital Entity(P2)

Digital Entity(Pn)

Storage

Storage

Storage

Internet

Application System

Archive Management System

Middleware

Development/ Collaboration Env.

Partnership

Discovery/ Federation

Archive Management

Portal

Archivist

Project Collaborators & Participants

Value-added Industry Partner

End Users
Digital Archive Grid Service Infrastructure

- User Requests
- HTML Data
- Digital Archive Portal
- Participant Node
- XML Data
- Participant Node
- Aggregated Data
- Service Metadata
- Detailed Object Data
- Service Metadata
- XML Data
- Service Metadata
- Detailed Object Data
- Participant Node
- Object Index Data
- Service Metadata
- Detailed Object Data
- Data Grid Nodes
- XML Data
- Service Metadata
- Detailed Object Data
Building Grid Service for DORE

- DORE (Document REtrieval) is
  - A middleware
  - A library
  - A tool
  - for programmers to develop metadata database applications

- DORE is a tool in Open Digital Archive Environment (ODAE).

- Migrate DORE applications to GT3 enabled, and also have backward compatibility to existing system.
Integrate existing DORE applications with GT3 front end

Client

interface

DOREGrid Factory

DOREGrid Service Instance

DOREGrid Service Instance

Existing DORE Applications (Academia Historica)

Existing DORE Applications (Academia Sinica)

Existing DORE Applications (National Palace Museum)
Plans for 2004

- Continue the development of GAT and services
  - more info in the following presentations
- More complicated scenarios
  - collaborative environments,
  - submitting, controlling and steering jobs from mobile devices,
  - more dynamic behavior of applications
- Two more GridLab Workshop Meetings (Lecce: 16-22 May, Zakopane: early December)
- Organizing the GridLab/GT3.2/WSRF integration meeting with US partners (to ensure GAT compatibility)
- Prepare for the Supercomputing demos
- GGF BOF (GAT) and finally GGF WG
- GGF Scheduling Architecture Working Group now started (GRMS)

Exploitation of the project results

- Close work with GridLab’s commercial partners
- Global Grid Application Alliance with GridLab’s leadership
- GGF activities (long term)
- GridSuite (based on GridLab plus Gridstart Open Source, PSNC plus partners)
  - Open Source + commercial support
- European Grid Support Centre (PSNC)

Visit to AIST, Tokyo, Japan 23 April, 2004