The 6th **COE Postdoctoral and Doctoral Researchers Technical Presentation**

A method of Data Management and Data Access on Wide Area Distributed Environment

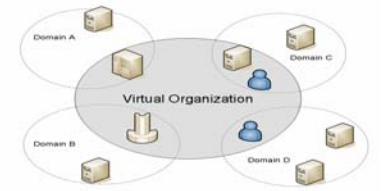
Minoru IKEBE (D1)
Laboratory for Internet Architecture and Systems
minoru-i@is.naist.jp

Background

- Various data dispersed existing in the Internet
  - User can access to data using search engine (such as Google), which publishing by WWW
- DataGrid Technology
  - DataGrid aims to manage a huge amount of data generated widely on the Internet
  - DataGrid mainly target
    - Scientific Simulation
      - Life Science / Bio Science
      - High Energy Physics / Astronomy
      - simulation program generate data more than 2 Terabyte per year
      - Executing Simulation program on Virtual Organization
    - Ubiquitous Network Sensors
      - Weather Sensors
      - Surveillance cameras
      - sensors generate a huge amount of data continuously

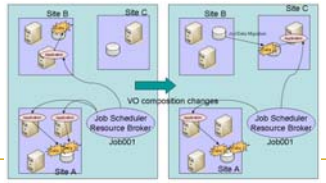
Virtual Organization (1)

- Virtual Organization (VO)
  - Dynamic collection of individuals and institutions which are required to share resources to achieve certain goals



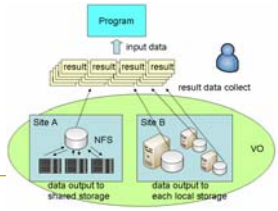
Virtual Organization (2)

- VO composition dynamic changes
  - VO comprise the available computing resource at that time
  - if each site users begin to use computing resources or resource break down
    - computer resource dynamic changes in VO



Result data

- Calculation result data
  - each computer output data to local storage or shared storage
  - user collect calculation results
  - user use the result for the other program

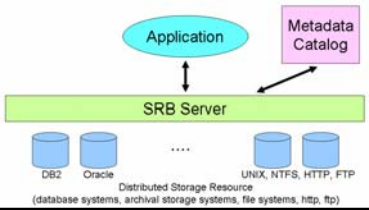


Users Requirements

- Users Requirements
  - access simulation data during and after executing simulation program
    - user changes parameters of simulation program while watching progress on the way
  - without being aware of the location of such data
    - during job executing
      - result data output to computing node's storage
      - user collect output data in that instant (snapshot data)
  - access result data by some semantic manner
    - It is difficult to one by one appoint data
    - semantic manner
      - simulation purpose
      - job submission time
      - and so on

Related Works

- Storage Resource Broker (SRB)
  - SRB is a Middleware for data management.
  - SRB was developed by the San Diego Supercomputing Center (SDSC).
  - It virtualizes resource access.
  - It mediates access to distributed heterogeneous resource.
  - It use Metadata to facilitate the brokering
  - It can't express Semantic manner



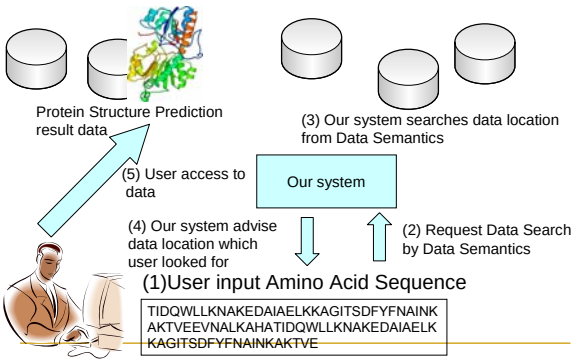
Our Research Purpose

- Our system provides the following functions for users
  - User can access data using semantic manner
  - Seamless access to data stored at dispersed sites
    - Seamless access means "without being aware of the location of such data"

Proposal System (1)

- System express semantic manner by Metadata
  - Metadata describes the attribute information about the data
- System provided Data Location Transparency for users
  - Metadata includes Data location information
- User search / access data using Metadata

Proposal System (2)

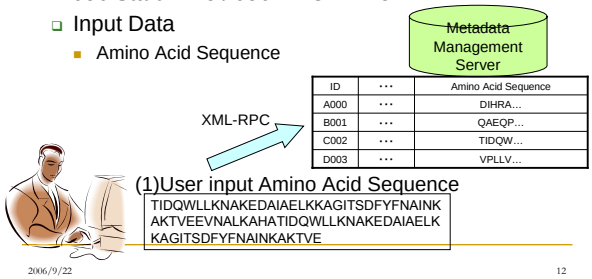


Metadata

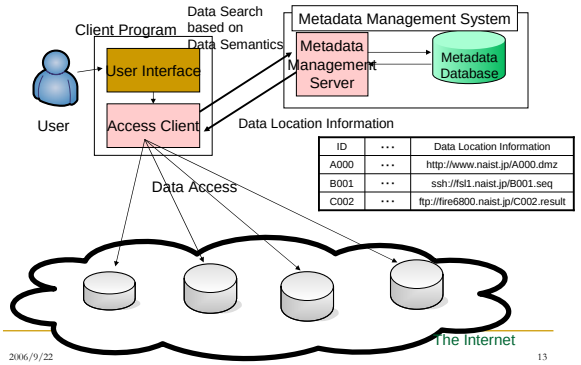
- Metadata represent Data Semantics
  - Metadata describes the attribute information about the data
  - Metadata describes to XML
  - System define two kind of Metadata
    - Basic Metadata
    - Application Metadata
- Basic Metadata
  - Common file attribute information
  - follow the Dublin Core
    - The Dublin Core metadata element set is a standard
- Application Metadata
  - application specific elements
  - define application metadata every application
  - In the case of "Predicts Protein 3D-structure"
    - input data (amino acid sequence)
    - job execute start / end time

Data Search based on Data Semantics

- Semantics Manner
  - Job Start Time / Job Finish Time
  - Input Data
    - Amino Acid Sequence



Data Access using Metadata



Making method of Metadata

- Basic Metadata
  - When data are generated
    - automatically obtain metadata elements
    - It is difficult to obtain the attribute information for all elements of Dublin Core Metadata
    - system can obtain a necessary Metadata element to access data
- Application Metadata
  - our system obtains application specify information from the Portal site

Future Plan (1)

- We have a plan to deploy our system on the PlanetLab.
  - PlanetLab is an open platform for developing, and accessing planetary-scale services
  - Current distribution of 701 nodes over 338 sites.

Future Plan (2)

- Devise Data Access / Search User Interfaces
  - Provides Data Reference Space for Users
    - If a user search for one keyword then every search result is the same result.
    - present high use data for users.