Several Issues of Information Retrieval in Grid Computing

Yaohui Ding
CS757 LDCSEE
West Virginia University

Overview

• Background and motivation
• Basic concepts
• Retrieval of data and information
• Conclusion
Background

- Increasingly large, complex and resource-intensive applications leads to a higher demand of computational resources.
- Within the computational resources, the development of storage and computing resources is unbalanced:
  - Microprocessors are reaching impressive speeds, in the long run they are falling behind storage.
  - The power of network develop even faster than the storage.
- One solution is to aggregate distributed resources in various collaborating institutions and to gather enough computational resources.
Grid Computing

-- Grid Computing Concept by Adarsh, from adarshpatil.com
Basic Concepts

- **Grid**: likened to the electrical power grid (appealing in 1960s)
  - Standard as plugging in an appliance into an outlet
- **Grid problem**: coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations [1]
  - **Resource**: including data, computers, scientific instruments, software, etc
  - Sharing must be *coordinated*
  - This coordination must be in a *dynamic* and *flexible* fashion
  - **Virtual organization (VO)**: a set of participants with various relationships that wish to share resources to perform some task

Overview

• Background and motivation
• Basic concepts
• Retrieval of data and information
• Conclusion
Grid Challenges and Technologies

- Security
- Uniform Access
- Computational Economy
- System Management
- Data locality
- Resource Discovery
- Resource Allocation & Scheduling
- Application Construction

Grid Computing, B. Wilkinson, 2004
Retrieving Data and Information

- Two types of Grid information
  - Information of *data* locations: e.g. find the file ‘foo’
  - Grid *resources*: e.g. find a CPU with some desired clock-rate

- Two models of information dissemination and discovery
  - Subscription-based dissemination
  - Query-based discovery

- Two types of architecture
  - Centralized services in the early years
  - Decentralized approach
    - with index servers and a registration protocol
    - Peer-to-peer
Discovering Data

• The goal is to find a physical data file given a logical file name
• Previous work in p-p networks can conceivably be utilized in Data Grid
  o e.g. CAN, Chord, Pastry and Tapestry

❖ Issue 1: Are there any characteristics of Grid VOs that can be exploited for optimizing data discovery?
  o Clue: The data is not produced and consumed by largely unrelated individuals
Discovering Data

- **Issue 1:** Are there any characteristics of Grid VOs that can be exploited for optimizing data discovery?

- **Possible Solution:**
  - The data access patterns result from specific scientific activities, not unrelated individuals (e.g. the Web)
  - Scientific communities sharing data tend to behave like “small-world networks”, which has 2 fundamental characteristics:
    - A small average path length
    - A loosely connected network of almost fully connected sub-networks
Discovering Data

- **Issue 2:** What data discovery scheme can take advantage of the small-world characteristics?

- **Possible Solution:**
  - Assuming the small-world topology is known, then
  - Data location information is propagated aggressively within clusters, while inter-cluster search uses request-forwarding techniques
Discovering Data

- **Issue 3:** What protocols will lead to a self-organizing overlay network that reflects the small-world properties of a VO?

- **Possible Solution:**
  - At that time, it is an open question
  - Recent work for this issue could be found in [2] [3]


Discovering Grid Resources

- In data discovery, every data item is assigned a unique identifier that is used for indexing and routing.
- In Grid resource discovery, services need to answer requests that specify desired sets of attribute values.

**Issue 4:** What request-forwarding scheme is appropriate for resource discovery on the Grid?
  - Clue: a single attribute (like a logical file name) is enough or not?
Discovering Grid Resources

- **Issue 4:** What request-forwarding scheme is appropriate for resource discovery on the Grid?

- **Possible Solution:**
  - [4] offers 4 simple request-forwarding strategies which allow peers to forward requests based on past experience.
  - The evaluation should consider some factors:
    - how many resources are managed by a peer?
    - how common is a resource?
    - the distribution of queries

Discovering Grid Resources

- **Issue 5:** What is the appropriate data model, and associated query language, for enabling Grid resource discovery?

- **Possible Solution:**
  1. Hierarchical model:
     - Performance is arguably sufficient
  2. Relational model:
     - More difficult to implement and scale
Discovering Grid Resources

- **Issue 6:** Regarding the Grid resources information, is consistency still a critical problem?

- **Possible Solution:**
  - Most Grid resource information follows a one-writer many-readers model
Replication

• Data replication is used in various distributed storage systems for improving performance and availability

• Here we focus on replicating application data in the Data Grid

  ❖ Issue 7: How does one ensure consistency among diverging replicas in a scalable way?
    ○ Clue: think about the specific nature of scientific communities
Replication

- **Issue 7:** How does one ensure consistency among diverging replicas in a scalable way?

- **Possible Solution:**
  - Current work does not place a large emphasis on this problem
  - In most scientific collaborations
    - Data is generated from instruments or derived from measured data
    - Then it is annotated with domain-specific metadata and published to the community
  - It is not true for all foreseeable applications, therefore, the issue will require increasing attention
Issue 8: What are efficient algorithms for creating and disseminating replicas in a Data Grid?

Possible Solution:

- [5] offers 5 caching/replication strategies inspired by the Web caching literature
  - Replicas are created according to changes in data access patterns
  - Access patterns of different scientific community should be better understand

Summary

• Grid computing is broad in its domain of application and raises research questions that span many areas of distributed computing
• Crucial features of the nature of scientific collaborations can be exploited to design appropriate distributed protocols and algorithms
• The discussion about aforementioned issues can offer some clue to the research of other new distributed computing approaches
Thank You!