Tutorial title: An Introduction to Data Grid Management Systems (DGMS)
Content Level: Introductory to Intermediate (New Topic/Area Introduction)
Length: Half day

Abstract

We describe a "grid" as a coordinated infrastructure, formed by combining resources that might be owned by distributed, autonomous administrative domains. A data grid infrastructure facilitates a logical view of distributed resources that are shared between autonomous administrative domains. An emerging problem in data storage industry is the management of unstructured data storage resources for inter/intra/multi-enterprise collaborative efforts. A new paradigm in data management systems, apart from traditional file systems and database systems is required to manage very large and distributed unstructured data. Data grids are being built around the world for coordinated sharing and management of unstructured data storage resources. The underlying concept behind data grids is the same as the concept behind relational databases: to isolate physical organization of the data from logical schema. In addition to isolating the logical view of unstructured data from the physical organization, data grids provide a logical view of different resources in the grid. This combined simple logical view of heterogeneous data and storage resources provides the flexibility needed to manage the very large unstructured data that is distributed in many enterprises.

The tutorial's objective is to introduce data grid technologies and their relevance to researchers and practitioners. The opportunities and challenges in this emerging field will be provided with reference to real projects. Novices and distributed data management experts would be benefited from this tutorial. An overview of a Data Grid Management System (DGMS) that manages more than a Petabyte of data would be provided. The tutorial would cover introduction, use-cases in large projects, design philosophies, existing technologies, open research issues, and demonstrations if possible.

Tutorial Outline

(***Note:*** *A detailed outline can be provided on request.*):

1. Introduction to DGMSs (**What?**) [5 minutes]
2. Example use cases of Data Grid Infrastructures deployed (**Where?)** [15 minutes]
3. Generic requirements for Data Grids (**Why?**) [20 minutes]
4. Design philosophies in DGMS (**How?**) [45 minutes]
5. DGMS Implementation Architecture – explained using SRB **(Working Example?)** [45 minutes]
6. Gridflows (Analysis pipelines) [20 minutes]
7. Related Technologies and Efforts [20 minutes]
8. Let us build a data grid in this room (Hands on for participants) [15 minutes]
9. Open Q&A session [25 minutes]

Speaker Profile

Arun swaran Jagatheesan *("Arun") is an OPS faculty member at the University of Florida, and a Visiting Scholar at the San Diego Supercomputer Center (SDSC) at University of California, San Diego. His research interests include Data Grid Management, Peer-to-peer Computing, and Workflow Management Systems. He is the founder and technical lead of the SRB Matrix Project on Gridflow Management Systems. He is a co-chair of the Grid File System Working Group at the Global Grid Forum. Arun is involved in research and development of multiple data grid projects at the San Diego Supercomputer Center.*

Equipment Requirements:

- Overhead Projector that can be connected to laptop
- If possible: High-speed Internet Connection to the Laptop

Expected Audience background:

Since the tutorial covers basics, existing production systems and open research issues, a wide variety of people usually fall into the category of "expected audience".

- *Beginners, Students:* Introduction on grid computing and data grids
- *Investigators:* Researchers from different computer science domains can update themselves on new research challenges in DGMS
- *System Managers and Consultants:* We already have multiple data grid projects in production now (with almost 1 Petabyte of data). These will be useful for project managers and consultants
- *Commercial companies: C*ase studies provide an idea of how similar problems in the commercial world could be solved and applied to handle collaborative data environments in inter/intra-organizational data management infrastructures.