Grid
MVP Guest Lecture

Henrik Thostrup Jensen

April 18th 2006

Outline

- What is grid, why is it needed
    - Visions, definitions, and challenges
- Grid Services
    - Grid as services, service types
- Grid Software
    - The Advanced Resource Connecter

What is Grid
Grid Services
Grid Software
Summary

Visions & Definitions
What is Grid Needed For
Challenges

What is Grid 1/4: The Grand Vision

A Huge Virtual Distributed Supercomputer

- Most agree that this cannot really work

What is Grid
Grid Services
Grid Software
Summary

Visions & Definitions
What is Grid Needed For
Challenges

What is Grid 2/4: Definition 1

A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.

The Grid - A Blueprint for a new Computing Infrastructure, 1998

What is Grid
Grid Services
Grid Software
Summary

Visions & Definitions
What is Grid Needed For
Challenges

What is Grid 3/4: Definition 2

The real and specific problem that underlies the Grid concept is coordinated resource sharing and problem solving in dynamic, multiinstitutional virtual organizations. The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem solving and resource-brokering strategies emerging in industry, science, and engineering. This sharing is, necessarily, highly controlled, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing occurs. A set of individuals and/or institutions defined by such sharing rules form what we call a virtual organization.

The anatomy of the Grid, 2000

What is Grid
Grid Services
Grid Software
Summary

Visions & Definitions
What is Grid Needed For
Challenges

What is the Grid 4/4: The Checklist

- Grid is a system that. . .
  - coordinates resources that are not subject to centralized control
  - using standard, open, general-purpose protocols and interfaces
  - to deliver nontrivial qualities of service.

- Grid is also a buzzword
  - Lots of software marked grid for PR purposes

What is Grid
Grid Services
Grid Software
Summary

Visions & Definitions
What is Grid Needed For
Challenges

What is Grid Needed For

- Sharing resources → better resource usage
  - Many resources spend a lot of time idle
  - Standard mechanisms for accessing resources
- E-science
  - Distributed collaboration, huge data sets
    - The LHC is the prime example
  - Many experiments are too big for a single organization
  - Many sciences completely depend on computers

What is Grid
Grid Services
Grid Software
Summary

Visions & Definitions
What is Grid Needed For
Challenges

Challenges in Grid

- Make hardware accessible to persons outside the organization that own the hardware
  - While continuing normal operation
  - Can control who gets access to what
  - And charge for use

- Authentication and authorization

- Protection of resources

- Control over resources

- Accounting

- Anonymity

What is Grid
Grid Services
Grid Software
Summary

Visions & Definitions
What is Grid Needed For
Challenges

Break

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

Grid as Services

- Resources are accessed through services
  - SOA has hit the grid (or vice versa)
  - Note: SOA $\gg$ Web Services
- A grid is composed of different services
- Grid service types
  - Job Execution, Data Storage, Information System
  - More exist, but these are the foundation

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

# Grid Services 1/4: Job Execution

- Most fundamental grid service
- Submit a job description, service executes job
- Usually a cluster or similar HPC equipment
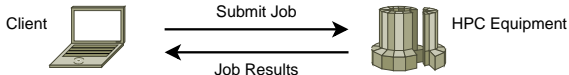
- Underlying details are not abstracted away
  - Grid is not about making resources look the same
- Most grids have their own submission interface
- Several job description languages exist

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

# Grid Services 2/4: Data Storage

- Upload data, retrieve it later
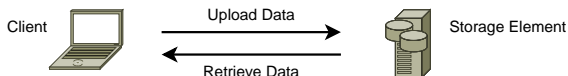
- Data often live for long periods of time
    - This makes it difficult to manage
- GridFTP is the common transfer protocol
    - Supports third party transfer
- HTTP, and others are also used

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

Grid Services 3/4: Information System

- Describes services - The backbone of a grid

- Aggregated Information Service
  - Provides high level information on other services
  - This is called resource discovery

Information System

Resource Discovery   Resource Registration

Client   Query Resource   Resource
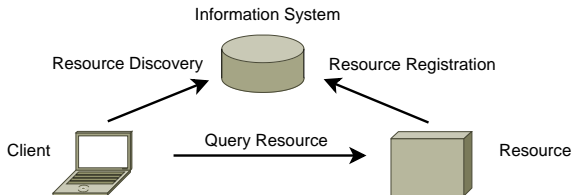
- Resource Information Service
  - Runs alongside every service and describes it

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

Grid Services 4/4

- Many service types exist
- Data Catalog
    - Keeping track of files
- Federated services
    - A service providing an abstraction over other service(s)
    - Examples: Work flow, data replication

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

Service Communication

- Services are not islands - they communicate
- Basic example
    - A service registers itself to the information system
    - Makes it possible for clients to discover them
- Job submission example
    - Input Data is staged in before job execution
    - Output files are stored on external storage
    - Output files are registered in data catalog
- This is done by the execution service
    - On behalf of the user

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

Grid Security

- All these services require secure access
- Requirements
  - Must work between administrative domains
  - Must be able to delegate rights
- Most existing security solutions are not usable
- A new security infrastructure have been created
  - GSI - Grid Security Infrastructure

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

Grid Security Infrastructure

- Based on Public Key Infrastructure (PKI)
- Host certificates work as usually in PKI
- Users are equipped with certificates as well
  - Primary certificate is not used directly
  - Proxy certificates are used instead
  - Proxy certificates enable delegation

What is Grid
Grid Services
Grid Software
Summary

Service Types
Service Communication
Grid Security

Break

What is Grid
Grid Services
Grid Software
Summary

Globus
ARC
ARC Services
ARC Advanced Topics

Grid Software

- Lots of half pieces have been constructed
  - We don't really know how the puzzle should look
- Grid software is immensely complex
- Many grid projects are now case studies in how not to use research money
- Some have been quite successful

What is Grid
Grid Services
Grid Software
Summary

Globus
ARC
ARC Services
ARC Advanced Topics

The Globus Toolkit

- The defacto grid middleware
- Lot of pioneering work
  - Have changed direction many times
  - Currently in version 4 (WSRF)
- Big - 95 MB of source code
  - Can be split up though
- Smaller projects are taking off
- Lot of different software is starting to talk together

What is Grid
Grid Services
Grid Software
Summary

Globus
ARC
ARC Services
ARC Advanced Topics

ARC: Advanced Resource Connecter

- Middleware developed in the Nordic countries
  - Based on Globus 2
- Runs a grid infrastructure today
  - About 60 sites, most are in Scandinavia
- Reasonably non-intrusive, relatively simple design
  - Only needs to be installed on the cluster front end
- Services:
  - Job execution, data storage, information system
  - Integrate with other services as well

What is Grid
Grid Services
Grid Software
Summary

Globus
ARC
ARC Services
ARC Advanced Topics

ARC Job Execution

- Submission happens over GridFTP
  - A joke/hack living on its fifth year
  - But works reasonably well
- The submission backend work with several LRMSes
  - PBS, SGE, Condor, LSF, Fork
  - Many grid middlewares only work with one

What is Grid
Grid Services
Grid Software
Summary

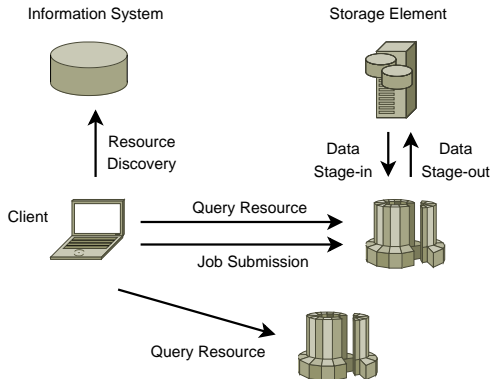Globus
ARC
ARC Services
ARC Advanced Topics

ARC Data Storage

- GridFTP and HTTP server available
  - HTTP server is close to unused
- GridFTP can use Posix mappings or GACL
- Does what is expected

What is Grid
Grid Services
Grid Software
Summary

Globus
ARC
ARC Services
ARC Advanced Topics

ARC Information System

- A Grid host has an information system describing it services
- Register its presence to index nodes
  - Several index nodes exists for redundancy
- Users query index nodes to discover resources

- Based on Globus MDS2
  - LDAP in disguise
- Schema that can describe clusters and storage elements
  - Services must have a schema to be described

What is Grid
Grid Services
Grid Software
Summary

Globus
ARC
ARC Services
ARC Advanced Topics

Job Submission Flow

ARC Brokering

- In ARC, brokering happens on the client side
  - No central bottleneck, no global optimization

- The user must be authorized to use the cluster and the queue
- The clusters characteristics must match the requirements in the job description
  - CPU time, free disk space, installed software
- From all clusters that fulfill the criteria one is chosen semi-randomly.
  - Clusters are weighted proportional to the number of free CPUs
  - If no available CPUs exists, the job is submitted to the cluster with the lowest number queued jobs pr CPU

Summary

- Grid - An infrastructure for resource sharing
- Grid as Services
  - Grid is composed of services
  - Job execution, data storage, information system
- Grid Software
  - Globus
  - ARC: Advanced Resource Connector