Implementing Production Grids

William E.Johnson,
The NASA IPG Engineering Team, and
The DOE Science Grid Team

Grid Computing – Making the Global Infrastructure a Reality
Fran Berman, Geoffrey Fox, and Tony Hey
2002 John Wiley & Sons, Ltd.
Chapter 5

Based on slides by
Kyung-Lang Park, Yonsei Univ. Supercomputing Lab.

Contents

Introduction

- There are a number of projects
  - UK's e-Science program, European DataGrid
  - Information Power Grid, DOE Science Grid
  - Asia-Pacific Grid, Ninf, GridLab
- They deploy a specific set of software
  - In the case of IPG & DOE Science Grid :
    - Globus, Condor, SRB/MCAT, PBSPro, and PKI authentication

Introduction

- Why the Globus package was chosen
  - A clear, strong, and standards based security model
  - Modular functions providing all of the Grid Common Services, except general events
  - A clear model for maintaining local control of resources that are incorporated into a Globus Grid
  - A general design approach that allows a decentralized control and deployment of the software
  - A demonstrated ability to accomplish large-scale Metacomputing
    - in particular SP-Express application in Gusto test bed
  - Presence in supercomputing environments
  - A clear commitment to open source
  - Today, one would also have to add "market share"

Introduction

- Why the Globus package was chosen (cont.)

- Legion and UNICORE were also considered, but failed to meet one or more of the selection criteria
- SRB and Condor were added because they provided specific, required functionality
- NASA promoted integration of these with Globus

The Grid Context

"Grids" are an approach for building dynamically constructed problem solving environments using geographically and organizationally dispersed, high performance computing and data handling resources.

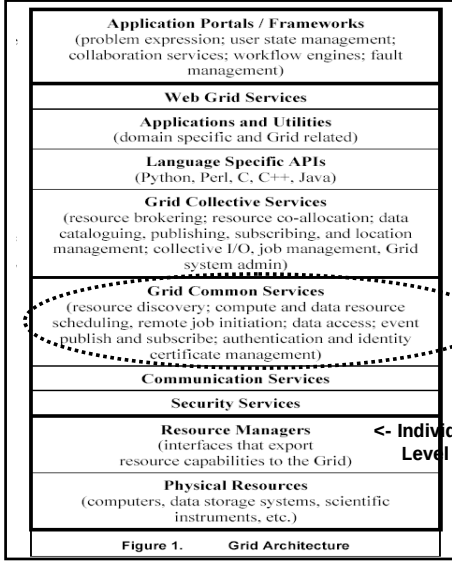
"Grids" also provide important infrastructure supporting multi-institutional collaboration

The Grid Context

- "Grids" are tools, middleware, and services for:
  - Building the application frameworks that allow discipline scientists to express and manage the simulation, analysis, and data management aspects of overall problem solving
  - Providing a uniform and secure access to a wide variety of distributed computing and data resources
  - Supporting construction, management, and use of widely distributed application systems
  - Facilitating human collaboration through common security services, and resource and data sharing
  - Providing support for remote access to, and operation of, scientific and engineering instrumentation systems
  - Managing and operating this computing and data infrastructure as a persistent service

The Grid Context

- This is accomplished through two aspects
  - A set of uniform software services
  - A widely deployed infrastructure



The software architecture of a Grid

The Grid Context

- Basic functions in order to be called a Grid (Grid Common Service, neck of the hourglass)
  - Grid Information Service (GIS) – resource discovery
  - Grid Security Infrastructure (GSI) -
  - Grid job initiator mechanism
  - Grid scheduling function
  - Basic data management (such as GridFTP)
  - In addition a Grid event mechanism is probably required:
    - Such as Grid Monitor Architecture (GMA)
  - A communication abstraction such as Globus I/O

Anticipated Grid Usage Model will determine what gets deployed, and when

- Grid computing models – vary from single resource to tightly coupled systems
  - Export existing services
  - Loosely coupled processes
  - Workflow managed processes
  - Distributed-pipelined/Coupled processes
  - Tightly coupled processes
- Grid data models
  - Occasional access to multiple tertiary storage systems
  - Distributed analysis of massive datasets
  - Large reference data sets
  - Grid metadata management

Grid Computing Models

- Export existing services
  - A uniform set of services export the capabilities of existing computing facilities
  - Accomplished by the Globus S/W, GridFTP
  - **User constructed system**
    - portals or framework that run on the user systems and provide for creating and managing related suites of Grid jobs
  - Grid compatible data service is also needed

Grid Computing Models

- Loosely Coupled Processes
  - Collections of logically related jobs that never-the-less do not have much in common once they are executing
  - E.g., Data analysis, parameter studies
  - Sometimes multiple input/output should be integrated into a single analysis or another input
  - **Job manager** is required
    - To track these related jobs in order to ensure either that they have all run exactly once, or that an accurate record is provided of those that ran and those that failed
  - E.g., Condor-G
    - Condor_manager, Blobus GASS server on client side
  - This is also the job model being addressed by "peer-to-peer" systems

Grid Computing Models

- Workflow Managed Processes
  - Existing application system frameworks have adhoc **workflow management element**
    - Manage events (asynchronous messages used for decision making purposes)
  - Most workflow managers handles "events" of all sorts
  - Grid events include:
    - Normal application occurrences
    - Abnormal application occurrences
    - Messages that certain data files have been written & closed
  - GMA defines an event model etc.

Grid Computing Models

- Distributed-pipelined processes
  - Multidisciplinary or other multi-component simulations will need to be executed in a "pipeline" fashion
  - **Co-scheduling** is likely to be essential
    - Scheduling multiple individual computing resources
    - E.g. at particular time-of-day
      - Supported by PBSPro, Maui scheduler
  - Advanced reservation scheduling

Grid Computing Models

- Tightly coupled processes
  - MPI and PVM style
  - Co-scheduling also needed
  - MPICH-G2 is a representative in Grid area
    - Handles co-scheduling
    - Uses manufactures MPI for local communication if available
    - May not work if other versions of MPICH are installed
  - PACX-MPI more mature but is not Grid services, because they do not make use of the Common Grid Services (e.g., Grid security services)
  - PVM can be under with Condor and Glide-in but does not use Grid authentication

Grid Data Models

- Occasional Access to multiple tertiary storage systems
  - Data mining can require access to metadata and uniform access to multiple data archives
  - SRB/MCAT
    - Uniform remote access, local caching, metadata catalogue for federation
  - GridFTP
    - Basic access capabilities as SRB for single data source
    - Intend to provide a standard, low-level Grid data access
  - GASS (Global Access to Secondary Storage)
    - Unix I/O style access to remote files
    - Copying the entire file to the local system and back on close
    - GASS put on servers near tertiary storage, and on user systems where input files managed

Grid Data Models

- Distributed Analysis of Massive Datasets Followed by Cataloguing and Archiving
  - Data intensive science disciplines
  - A replica catalogue and a replica manager, and a high performance data movement tool (GridFTP) are needed
  - GridFTP should provides:
    - Integrated GSI security and policy-based access
    - Third-party transfers
    - WAN parameter optimization
    - Partial fie access
    - Reliability/restart
    - Integrated performance monitoring
    - Network parallel transfer streams
    - Server side data striping/computation
    - Proxies

Grid Data Models

- Large reference data sets
  - Network cache 'close to' computational resources would be useful
  - Distributed Parallel Storage System (DPSS) but not well integrated with Globus
- Grid metadata management
  - MCAT/SRB supports
  - Heavyweight & requires considerable operational support

Grid Support for Collaboration

- Support for Virtual Organizations (VO)
- Grid Security Infrastructure
  - Common authentication approach
  - Essential aspect of collaboration
- Preserve and share the organization structure
  - GIS provide this service
- Access Grid
  - Video conferencing facility

Building an Initial Multi-site, Computational and Data Grid

- The Grid building team
  - Good working relationships is essential
- Grid resources
  - Identify the computing and storage resources to be incorporated into your Grid
  - Choose a batch scheduler, carefully consider the issue of co-scheduling

Building an Initial Multi-site, Computational and Data Grid

- Build the initial testbed
  - Grid information service (MDS)
    - To locate resource based on job characteristics (OS, # CPUs, memory, etc)
    - Grid Resource Information Service (GRIS)
    - Grid Information Index Server (GIIS) – user accessible, supports searching
    - Plan a GIIS at each site with significant resources to avoid single point of failure
    - May need several GIIS at site to handle search load
  - Build Globus on test systems
    - Use PKI authentication
    - Certificates from Globus Certificate Authority (CA)
    - Validate access to GIS/GIIS at all sites, local and remote job submission

Cross-site Trust Management

- Uniform Grid entity naming and authentication is important
- PKI, X.509, TLS/SSL are understood and largely accepted
- The real issue Certification Authority ("CA")
- In the PKI authentication environement the CA policies encoded in formal documents called *Certificate Policies/Certification Practice (CP)*

Cross-site Trust Management

- 6.1 Trust
  - Cyberspace trust begins with clear, transparent, negotiated, and documented policies associated with identity
  - When a Grid identity token (X.500 certificate) is verified, one should have this assurance
  - Policy associated with identity
    - Difficulty depends on nature of VO
      - Administratively similar systems (e.g. in organization)
        - » Informal/existing trust model can be extended
      - Administratively diverse systems (e.g NASA, DOE labs)
        - » Formal/existing trust model can be extended
      - Administratively heterogeneous
        - » Formal/new model required

6.2 Establishing an Operational CA

- Set up a CA to issue Grid X.509 identity certificates
  - DOE uses Netscape CMS
  - GGF working on standard set of CPs as templates
- 6.2.1 Naming
  - Use flat namespace
  - Do not try to encode all information in names
  - Do not authorize on basis of components of name
  - Problems
    - People belong to multiple organizations with different authorizations

6.2.2 The Certification Authority Model

- It is formal and expensive to operate CA
- Groups finding shared CA
- Model – central CA with overall CP
  - Subordinate CPs for collcetions of VOs
  - CA delegates to VOs issues of membership and subscription
  - Model used by DOE
- Root CA (locked-up and offline!!) signs certs of CAs that issue user certs
- Registration Managers (RMs) operated by VOs

Cross-site Trust Management

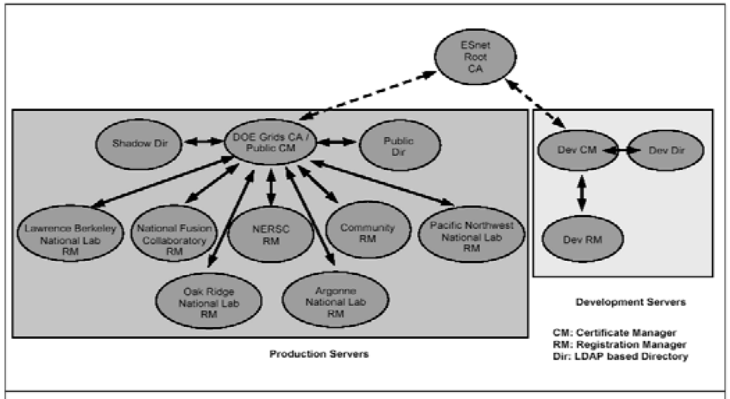


Figure 2. Software Architecture for 5/15/02 Deployment of the DOE Grids CA.
(Courtesy Tony Genovese (tony@es.net) and Mike Helm (helm@es.net). ESnet, Lawrence Berkeley National Laboratory.)

Transition to a Production Grid

- First steps
  - Issue host certificates for all the computing and data resources
  - establish procedure for installing them
  - Issue user certificates
  - Using certificates issued by your CA, validate correct operation of the Grid Security Infrastructure (GSI), GSS libs, GSISSH, GSIFTP or GridFTP at all sites

Transition to a Production Grid

- Defining/Understanding the Extent of "Your" Grid
  - Boundaries of a Grid are primarily determined by these factors
    - Interoperability of the Grid software
      - Most sites with Globus can
    - What CAs you trust
      - Configured in Globus environment
    - How you scope the searching of the GIS/GIIS

Transition to a Production Grid

- The model for the Grid Information System
  - Servers above local GIIS expand search space and provide large collection of resources transparently
  - Two options:
    - An X.500 style hierarchical name component space directory structure
    - Index server directory structure (MDS)

Transition to a Production Grid

- Local Authorization
  - Globus mapfile is an ACL that maps from Grid identities to local UIDs on the systems where jobs are to be run
  - Does not scale well
  - Can map multiple Grid users to single account
  - Future: CAS (Community Authorization Service)

Transition to a Production Grid

- Site security issues
  - IP communication ports – many used by Grid services
  - Firewall issue
  - Globus can be configured to use a restricted range of ports
    - Several 10s in mid-700s still needed
  - Need to ensure net security engineers do not close accidentally
  - The DOE Science Grid is in the process of defining Grid firewall policy document

Transition to a Production Grid

- High performance communication
  - Enlist the help of a WAN networking specialist
  - Check and refine the network bandwidth end-to-end using large packet size test
  - NetLogger, pipechar for debugging

Transition to a Production Grid

- Batch Scheduler
  - Provide good interface for Globus GRAM
  - PBS can provide time-of-day based advanced reservation
  - PBS can pass Grid ID to accounting system
  - PBS supports access-controla and high-priority queues
  - People attached to scheduler sysntax
    - PBS componentized to use PBS commands to submit Globus jobs

Transition to a Production Grid

- Preparing for Users
  - Try and find problems before your users do
  - Validation suites
- Moving from Testbed to prototype production Grid
  - Build Globus on 2 or maore production computing platforms at different sites and checks
  - Link Globus job submission with local batch schedulers
- Grid System Administration Tools
  - Grid monitoring tool
  - pyGlobus modules for NetSaint to test GSIFTP, MPS, and Globus gatekeeper

Transition to a Production Grid

- Data Management and Your Grid Service Model
  - Establish the model for moving data between all of the systems involved in your Grid
  - If user sites will manage data, GridFTP servers should be deployed there
    - Sysadm issue at such sites to manage Grid components

Transition to a Production Grid

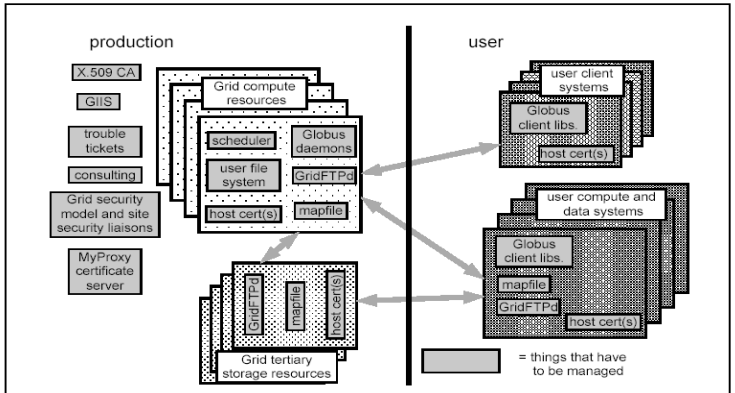


Figure 4. Establish Your Grid Service Model

Transition to a Production Grid

- Take Good Care of the Users as early as possible
  - If at all possible, establish a Grid/Globus application specialist group
  - Currently Grid use difficult due to low level primitives
  - high level framework will be better e.g. OGSA etc
  - Grid tracking/monitoring portal
    - IPG LauchPad
    - NPACI HotPage

Transition to a Production Grid

- MyProxy Service
  - Normal proxies have short life (12h)
  - myProxy simplifies user management of certificates
  - Provide for creating and storing intermediate lifetime proxies that may be accessed by on behalf of the user
  - Provides a set of client tools that let the user create, store, and destroy proxies
  - Needs to be secure persistent server

Conclusion

- Actual experience was presented in building two production Grids, IPG and DOE
- There is a lot more Grid softwarenow
  - But Globus, SRB, Condor still central

Reference

- Great reference section!!
  - Listing all sorts of Grid research and summarizing
- Maybe more valuable than main contents
- 108 materials are included
  - UK eScience, EU DataGrid, NASA IPG, DOE, AP Grid, Ninf, NetCFD, GridLab, nersc, globus, condor, srb, pbs, pki, esnet, virtual observatory, gis, gram, gridftp, GMA, DMF, globus i/o, maui, gridport, cactus, netsolve, Ilab, MPICH-G2, PVM, PACX-MPI, G┴┼, GASS, replica catalog, globus replica management, GDMP, Access Grid, Grid-Enabled Openssh, TCP tunning guide