

Implementing Production Grids

*William E. Johnston (wejohnston@lbl.gov),
The NASA IPG Engineering Team (www.ipg.nasa.gov), and
The DOE Science Grid Team (doesciencegrid.org)*
(These slides are available at grid.lbl.gov/~wej/Grids)

Architecture of a Grid

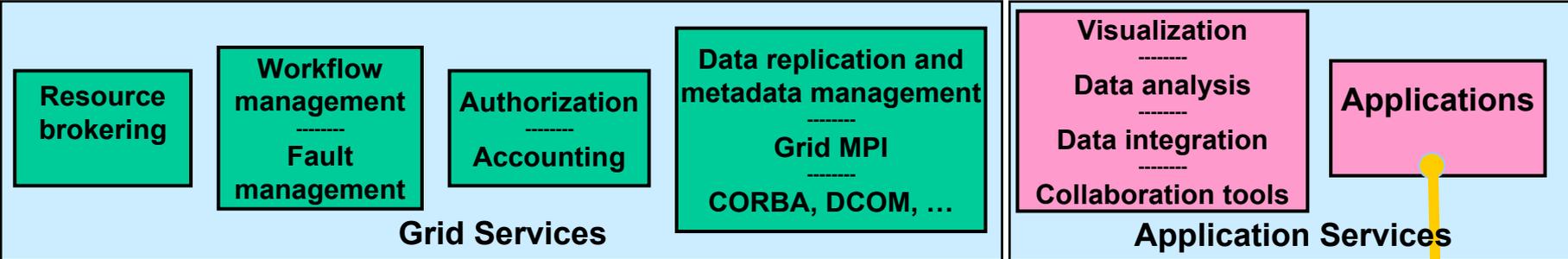


Portals

Portals that are Web Services based, shell scripts, specialized (e.g. high end vis workstations, PDAs)

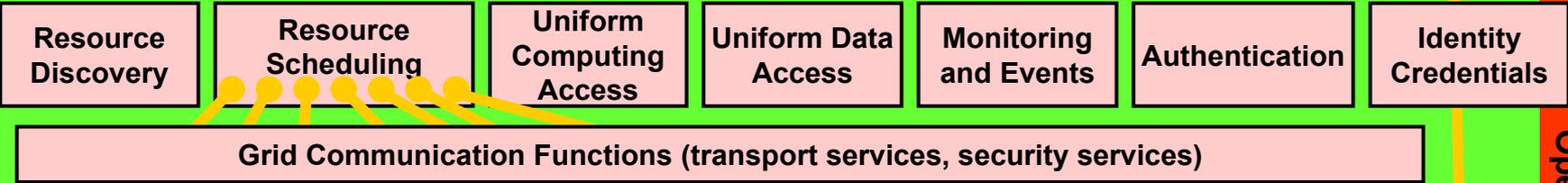
Encapsulation as Web Services, as Script Based Services, as Java Based Services

Advanced Services



Encapsulation as Web Services, as Script Based Services, as Java Based Services

Basic Grid Services



Communications



Distributed Resources



scientific instruments



clusters



Condor pools of workstations



tertiary storage

national supercomputer facilities



job initiation, event generators, GridFTP servers

Operational Support

Lessons Learned for Building Large-Scale Grids

- Deploying operational infrastructure
- Cross site trust
- Dealing with Grid technology scaling issues
- Listening to the users

The Anticipated Grid Usage Model Will Determine What Gets Deployed, and When

- Grid computing model (processes running on multiple systems)
 - Loosely coupled processes
 - data analysis
 - Condor-G manages groups of related jobs
 - workflow managed
 - event services will probably be needed to drive the workflow
 - Coupled processes
 - Pipelined
 - Co-scheduling will be needed
 - Tightly coupled processes
 - cross system MPI and co-scheduling

Grid Usage Models

- Grid data model
 - Occasional access to multiple tertiary storage systems
 - Data mining
 - GridFTP, SRB
 - Distributed analysis of massive datasets followed by cataloguing and archiving
 - high energy physics
 - DataGrid tools, SRB/MCAT
 - Large reference data sets
 - Must be available for a series of calculations, but are read-only
 - Reservable caches

Grid Usage Models

- Grid collaboration model
 - Across administratively similar systems
 - Within an organization
 - Informal / existing trust model extended to Grid authentication and authorization
 - Administratively diverse systems
 - Across many similar organizations (e.g. NASA Centers, DOE Labs)
 - Formal / existing trust model extended to Grid authentication and authorization
 - Administratively heterogeneous
 - Across multiple organizational types (e.g. science labs and industry)
 - International collaborations
 - Formal / new trust model for Grid authentication and authorization
 - The Certificate Policy and Certificate Practices Statement are critical

Building a Multi-site, Computational and Data Grid

- Like networking, successful Grids involve almost as much sociology as technology.
- The first step is to establish the mechanisms for promoting cooperation and mutual technical support among those who will build and manage the Grid.
- Establish an Engineering Working Group that involves the Grid deployment teams at each site
 - schedule weekly meetings / telecons
 - involve Globus experts in these meetings
 - establish an EngWG archived email list

The Grid Building Team

- Set up liaisons with the system administrators for all systems that will be involved (computation and storage)
 - this is especially important if the resources that you expect to incorporate in your Grid are
 - not in your organization
 - not in your part of your organization
 - Make sure everyone understands that:
 - Grid software involves not only root processes, but a different trust model for authorizing users
 - local control of resources is maintained, but managed differently

Grid Resources

- Identify the computing and storage resources to be incorporated into your Grid
 - be sensitive to the fact that opening up systems to Grid users may turn lightly or moderately loaded systems into heavily loaded systems
 - batch schedulers may have to be installed on systems that previously did not use them in order to manage the increased load
 - carefully consider the issue of co-scheduling!
 - many potential Grid applications need this
 - only a few available schedulers provide it (e.g. PBSPro)
 - this is an important issue for building distributed systems

Build the Initial Testbed

- Plan for a Grid Information Service / Grid Information Index Server (GIS/GIIS) at each distinct site with significant resources
 - this is important in order to avoid single points of failure
 - if you depend on an MDS/GIIS at some other site, and it becomes un-available, you will not be able to examine your local resources
- The initial testbed GIS/MDS model can be independent GIISs at each site
 - in this model
 - Either cross-site searches require explicit knowledge of each of the GIISs, which have to be searched independently, or
 - All resources cross-register in each GIIS

Build the Initial Testbed

- Build Globus on test systems
 - use PKI authentication and certificates from the Globus Certificate Authority, or some other CA, issued certificates for this test environment
 - can use the OpenSSL CA to issue your own certs manually
 - validate the access to, and operation of the GIS/GIISs at all sites

Preparing for the Transition to a Prototype-Production Grid

- There are a number of significant issues that have to be addressed before going to even a pseudo production Grid
 - Policy and mechanism must be established for the Grid X.509 identity certificates
 - the operational model for the Grid Information Service must be determined
 - who maintains the underlying data?
 - the model and mechanisms for user authorization must be established
 - how are the Grid mapfiles managed?
 - your Grid resource service model must be established (more later)
 - your Grid user support service model must be established
 - Documentation must be published

Trust Management

- Trust results from clear, transparent, and negotiated policies associated with identity
- The nature of the policy associated with identity certificates depends a great deal on the nature of your Grid community
 - It is relatively easy to establish policy for homogeneous communities as in a single organization
 - It is very hard to establish trust for large, heterogeneous virtual organizations involving people from multiple, international institutions

Trust Management

- Assuming a PKI Based Grid Security Infrastructure (GSI)
- Set up, or identify, a Certification Authority to issue Grid X.509 identity certificates to users and hosts (many use the Netscape CMS software for this)
- Make sure that you understand the issues associated the Certificate Policy / Certificate Practices (“CP”) of the CA
 - one thing governed by CP is the “nature” of identity verification needed to issue a certificate (this is a primary factor in determining who will be willing to accept your certificates as adequate authentication for resource access)
 - changing this aspect of the CP could well mean not just re-issuing all certificates, but requiring all users to re-apply for certificates

Trust Management

- Do not try and invent your own CP
- The GGF is working on a standard set of CPs
- The DOE Science Grid is working on supporting a multiplicity of international collaborations, and this is pushing the evolution of the GGF CP
- Establish and publish your Grid CP

PKI Based Grid Security Infrastructure (GSI)

- Pay very careful attention to the subject namespace
 - the X.509 Distinguished Name (the full form of the certificate subject name) is based on an X.500 style hierarchical namespace
 - if you put institutional names in certificates, don't use colloquial names for institutions - consider their full organizational hierarchy in defining the naming hierarchy
 - find out if anyone else in your institution, agency, university, etc., is working on PKI (most likely in the administrative or business units) - make sure that your names do not conflict with theirs, and if possible follow the same name hierarchy conventions
 - CAs set up by the business units of your organization frequently do not have the right policies to accommodate Grid users

PKI Based Grid Security Infrastructure (GSI)

- Increasingly there is an understanding that the less information of any kind put into a certificate the better
 - This simplifies certificate management and re-issuance when users forget passphrases
 - It emphasizes that all trust is established “locally” (by the resource owners and/or when joining a virtual community)
 - Potentially simplifies user management of certificates when that user is a member of multiple communities
 - For example, the CA run by ESNet for DOE will service several dozen different virtual communities
 - uses a flat namespace, with a random string of numerical digits to ensure name uniqueness
 - has separate Registration Authorities for each virtual community
 - has a Policy Management Authority to ensure an agreed upon minimum level of identity integrity is maintained among the communities
 - envisage.es.net

PKI Based Grid Security Infrastructure (GSI)

- Think carefully about the space of entities for which you will have to issue certificates
 - Humans
 - Hosts (systems)
 - Services (e.g. GridFTP)
 - Security domain gateways (e.g. PKI to Kerberos)
- Each must have a clear policy and procedure described in your CA's CP/CPS

Preparing for the Transition to a Prototype-Production Grid

- Issue host certificates for all the resources and establish procedures for installing them
- Count on revoking and re-issuing all of the certificates at least once before going operational
- Using certificates issued by your CA, validate correct operation of the GSI/GSS libraries, GSI ssh, and GSIftp / Gridftp at all sites

Defining / Understanding the Extent of “Your” Grid

- The “boundaries” of a Grid are primarily determined by three factors:
 - Interoperability of the Grid software
 - Many Grid sites run some variation of the Globus software, and there is fairly good interoperability between versions of Globus, so most Globus sites can potentially interoperate
 - what CAs you trust
 - this is explicitly configured in each Globus environment
 - how you scope the searching of the GIS/GIISs or control the information that is published in them
 - this depends on the model that you choose for structuring your directory services

Defining the Extent of “Your” Grid

- Your trusted CAs establishes the maximum extent of your user population
 - however there is no guarantee that every resource in what you think is “your” Grid trusts the same set of CAs – i.e. each resource potentially has a different space of users
 - in fact, this will be the norm if the resources are involved in multiple virtual organizations as they frequently are, e.g., in the high energy physics experiment data analysis communities

The Model for the Grid Information System

- There are currently two main approaches that are being considered for building directory services above the GIISs
- These higher level directories support
 - Expanding the resource search space through cross-GIIS searches for resources
 - Addressing scaling issues
 - Query caching, high-speed search engines, etc.
 - hosting / defining virtual organizations
 - hosting / providing views of collections of data objects that reside in different storage systems (federating)

The Model for the Grid Information System

- The two approaches are
 - ❖ a hierarchically structured set of directory servers and a managed namespace, al la DNS and X.500
 - ❖ “index” servers that provide views of a specific set of other servers, such as a collection of GIISs, data collections, etc.
- Both provide for “scoping” your Grid in terms of the resource search space

The Model for the Grid Information System

- An X.500 style hierarchical name component space
 - has the advantage of organizationally meaningful names that represent a set of “natural” boundaries for scoping searches
 - potentially can use commercial metadirectory servers for better scaling
 - attaching virtual orgs., data name spaces, etc. to the hierarchy makes them automatically visible, searchable, and in some sense “permanent” (because they are part of a managed name space)
 - try and involve someone who has some X.500 experience
 - don’t use colloquial names for institutions - consider their full organizational hierarchy when naming
 - many Grids use o=grid as the top level
 - Has the disadvantage that they are notoriously hard to get right, a situation that is compounded if Vos are included in the namespace

The Model for the Grid Information System

- Index servers
 - resources are typically named using the components of their DNS name
 - advantage is that of using an established and managed name space
 - must use separate “index” servers to define different relationships among GIISs, virtual organization, data collections, etc.
 - on the other hand, you can establish “arbitrary” relationships within the collection of indexed objects
 - this is the approach favored by the Globus R&D team

Local Authorization

- Establish the conventions for the Globus mapfile
 - maps user Grid identities to system UIDs – this is the basic local authorization mechanism for each individual platform, e.g. compute and storage
 - establish the connection between user accounts on individual platforms and requests for Globus access on those systems
 - if your Grid users are to be automatically given accounts on a lot of different systems, it may make sense to centrally manage the mapfile and periodically distribute it to all systems
 - however, unless the systems are administratively homogeneous, a non-intrusive mechanism such as email to the responsible sys admins to modify the mapfile is best
 - Community Authorization Service (CAS)

Site Security Issues

- Establish agreements on firewall issues
 - Globus can be configured to use a restricted range of ports, but it still needs several tens, or so (depending on the level of usage of the resources behind the firewall), in the mid 700s
 - A Globus “port catalogue” is available to tell what each Globus port is used for
 - this lets you provide information that you site security folks will likely want
 - should let you estimate how many ports have to be opened (how many per process, per resource, etc.)
 - GIS/MDS also needs some ports open
 - CA typically uses a secure Web interface (port 443)
- Develop tools/procedures to periodically check that the ports remain open

High Performance Communciation

- If you anticipate high data-rate distributed applications
 - enlist the help of a networking type and check and refine the network bandwidth end-to-end using large packet size data streams
 - problems are likely between application host and site LAN/WAN gateways, and along any path that traverses the commodity Internet
 - provide the application developers with end-to-end monitoring libraries/toolkits (e.g. Netlogger [see refs]) and tools like pipechar [see refs])
 - try to provide network monitors capable of monitoring specific TCP flows and returning that information to the application for the purposes of performance debugging

Preparing for Users

- Build and test your Grid incrementally
 - very early on, identify a test case distributed application that requires reasonable bandwidth, and run it across as many widely separated systems in your Grid as possible
 - try and find problems before your users do
 - design test and validation suites that exercise your Grid in the same way that applications do
- Establish user help mechanisms
 - Grid user email list and / or trouble ticket system
 - Web pages with pointers to documentation
 - a Globus “Quick Start Guide” that is modified to be specific to your Grid, with examples that will work in your environment (starting with a Grid “hello world” example)

The End of the Testbed Phase

- At this point Globus, the GIS/MDS, and the security infrastructure should all be operational on the testbed system(s). The Globus deployment team should be familiar with the install and operation issues, and the sys admins of the target resources should be engaged.
- Next step is to build a prototype-production environment.

Moving from Testbed to Prototype Production Grid

- Deploy and build Globus on at least two production computing platforms at two different sites. Establish the relationship between Globus job submission and the local batch schedulers (one queue, several queues, a Globus queue, etc.)
- Validate operation of this configuration

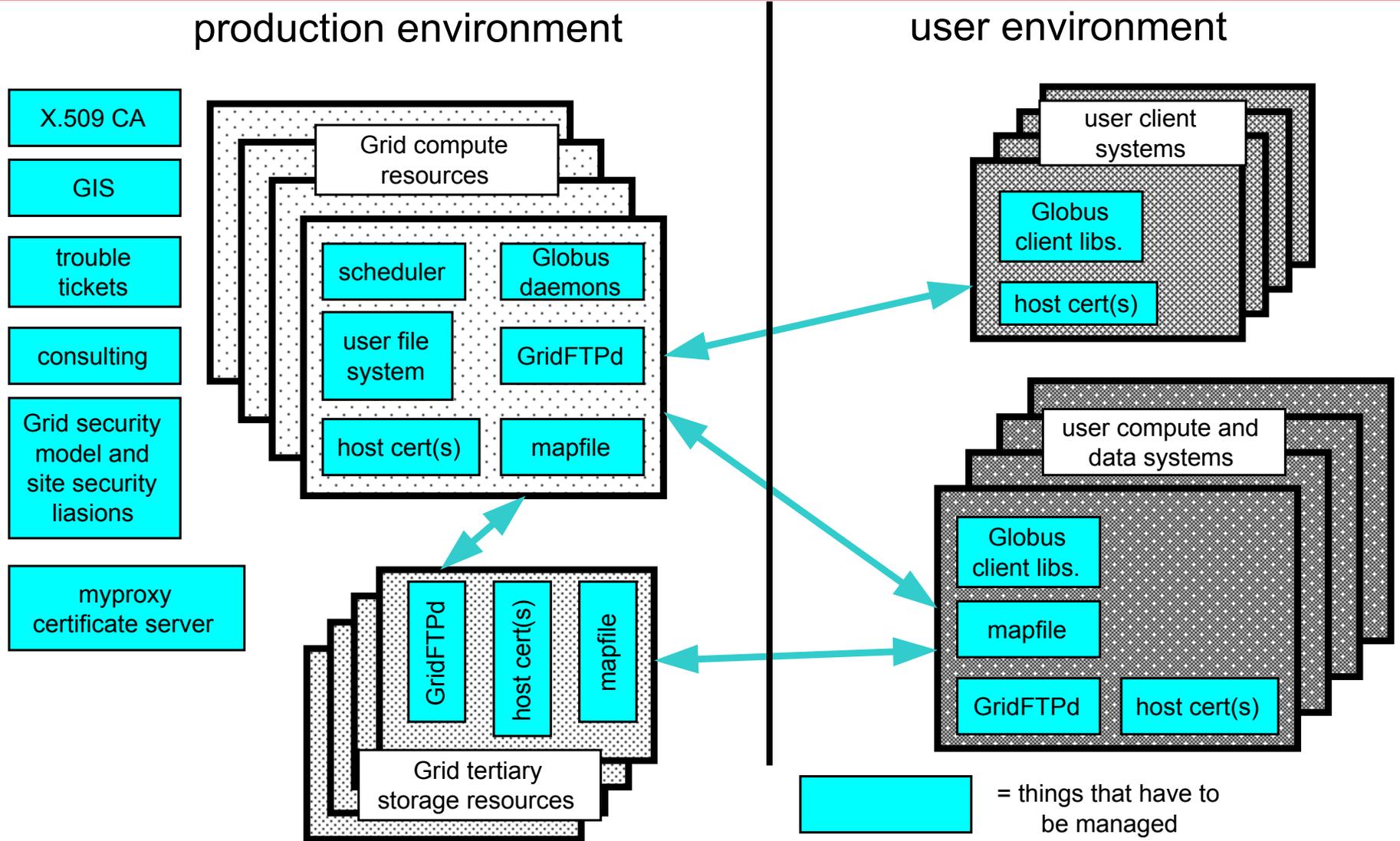
Data Management

- Establish the model for moving data between the Grid systems.
 - GSIFTP / GridFTP servers should be deployed on the Grid computing platforms and on the Grid data storage platforms
 - In later versions of Globus (1.1.4, and up) the restriction on forwarding of user proxies by third parties does not apply to the basic data services (GridFTP and GASS)
 - i.e., a job submitted from platform_1@site_A to platform_1@site_B to write back to a storage systems at site A (platform_2@site_A)

Data Management

- Establish the model for moving data between the all of the systems involved in your Grid
 - Determine if any user systems will manage user data that are to be used in Grid jobs.
 - This is common in the scientific environment where individual groups will manage their experiment data, e.g., on their own systems
 - If user systems will manage data, then the GridFTP server should be installed on those systems so that data may be moved from user system to user job on the computing platform, and back
 - Establish your service model
 - offering GridFTP on user systems may be essential, however managing long lived / root access Grid components on user systems may be “tricky” and/or require system admin work on your part
 - Validate that all of these data paths work correctly

Steps to Setting Up a Multi-Site Grid



Establish Your Grid Service Model

Take Good Care of the Users as Early as Possible

- ***Establish a Grid/Globus application specialist group***
 - they should be running sample jobs as soon as the testbed is stable, and certainly as soon as the prototype-production system is operational
 - they should serve as the interface between users and the Globus system administrators to solve Globus related application problems

- ***Identify early users and have the Grid/Globus application specialists assist them in getting jobs running on the Grid***
 - One of the scaling / impediment-to-use issues currently is that the Grid services are relatively primitive (I.e., at a low level). The Grid Services and Web Grid Services work currently in progress is trying to address this.

Take Good Care of the Users as Early as Possible

- Decide on a Grid job tracking and monitoring strategy
- Put up one of the various Web portals for Grid resource monitoring
- Consider using a myProxyServer to simplify user management of certificates