U.S. ATLAS

U.S. ATLAS Grid Perspectives

John Huth
Harvard University

U.S. ATLAS

Outline

- Tools in use in US ATLAS
- Production in US ATLAS
- Comparison of EDG/Nordic/US grids
- Comments on minimal set of tools required for DC1P2

U.S. ATLAS

Tools tested in US

- Common software: globus 2.0, VDT, Condor…
- Package distribution: pacman
- Metadata- file registry, production DB: Magda
- Job submission portal: Grappa
- Job control scripting: Grat
- Grid monitoring: gridview, ganglia
- Athena monitoring: netlogger
- Distributed analysis (development): dial

U.S. ATLAS

PACMAN

- PACkageMANager
- Configuring of automatic, simple setup of *complete* SW package for use on all grid sites
- Successfully employed for US ATLAS grid distribution, usage is growing to CMS, elsewhere
- It lets you define how a mixed tarball/rpm/gpt/native software environment is:

U.S. ATLAS

PACMAN Features

- Fetched
- Installed
- Setup
- Updated

`my_environment.pacman`

Experts can be figure this out once and exported to the rest of the world via trusted **caches**.

```
% pacman -get my_environment
```

Pacman Local Installation - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\My Documents\Atlas\Oxford\1.html

Pacman Local Installation

- Caches in Use -

- file:/software/install/
- Local Test Cache
- file:/home/youssef/sw/atlas/pacman/localcache/
- http://physics.bu.edu/~youssef/pacman/sample_cache/
- http://heppc12.uta.edu/pacman/
- http://www.usatlas.bnl.gov/computing/software/pacman/ACF-cache/
- http://iuatlas.physics.indiana.edu/pacman/IUcache/
- http://atlasgrid.bu.edu/atlasgrid/atlas/atlas_cache/

The caches you have decided to trust

Dependencies are automatically resolved

- Local Packages -

Package	Description	Installed?	Update?	Depends on	Daemons	Location	Date fetched
Pacman	Package Manager	yes	offline	-	-	-	Sun May 19 21:41:52 2002
atlas_testbed	Atlas Testbed Setup	yes	-	atlas-globus atlfast Pacman nedit	-	-	Sun May 19 21:41:52 2002
atlfast	Grid atlfast	yes	-	atlfast-boxed-0.2	-	/software/install	Sun May 19 21:41:43 2002
atlfast-boxed-0.2	Standalone grid atlfast	yes	-	-	-	/software/install	Sun May 19 21:41:43 2002
nedit	Text Editor	native	-	-	-	-	Sun May 19 21:41:52 2002

Installation as of Tue May 21 08:53:11 2002.

Pacman Home

Installed software, pointer to local documentation

My Computer

U.S. ATLAS

PACMAN USERS

VDT, the virtual data toolkit

The Atlas testbed environment (http://atlasgrid.bu.edu/atlasgrid/atlas/testbed) used for DC1 production.

Pacmanization of the CMS testbed is ongoing.

There have been more than 5000 downloads of the software

Developed at Boston University, support and further development are in collaboration with the Condor/VDT team at Wisconsin.

U.S. ATLAS

MAGDA

- Manager for Grid-based Data
- Prototype for distributed data management
  - Incorporates robust grid tools as they become available
- Uses MySQL, PERL, JAVA, and C++
- In use by US ATLAS grid testbed for file registration, transfers from DC1 tests

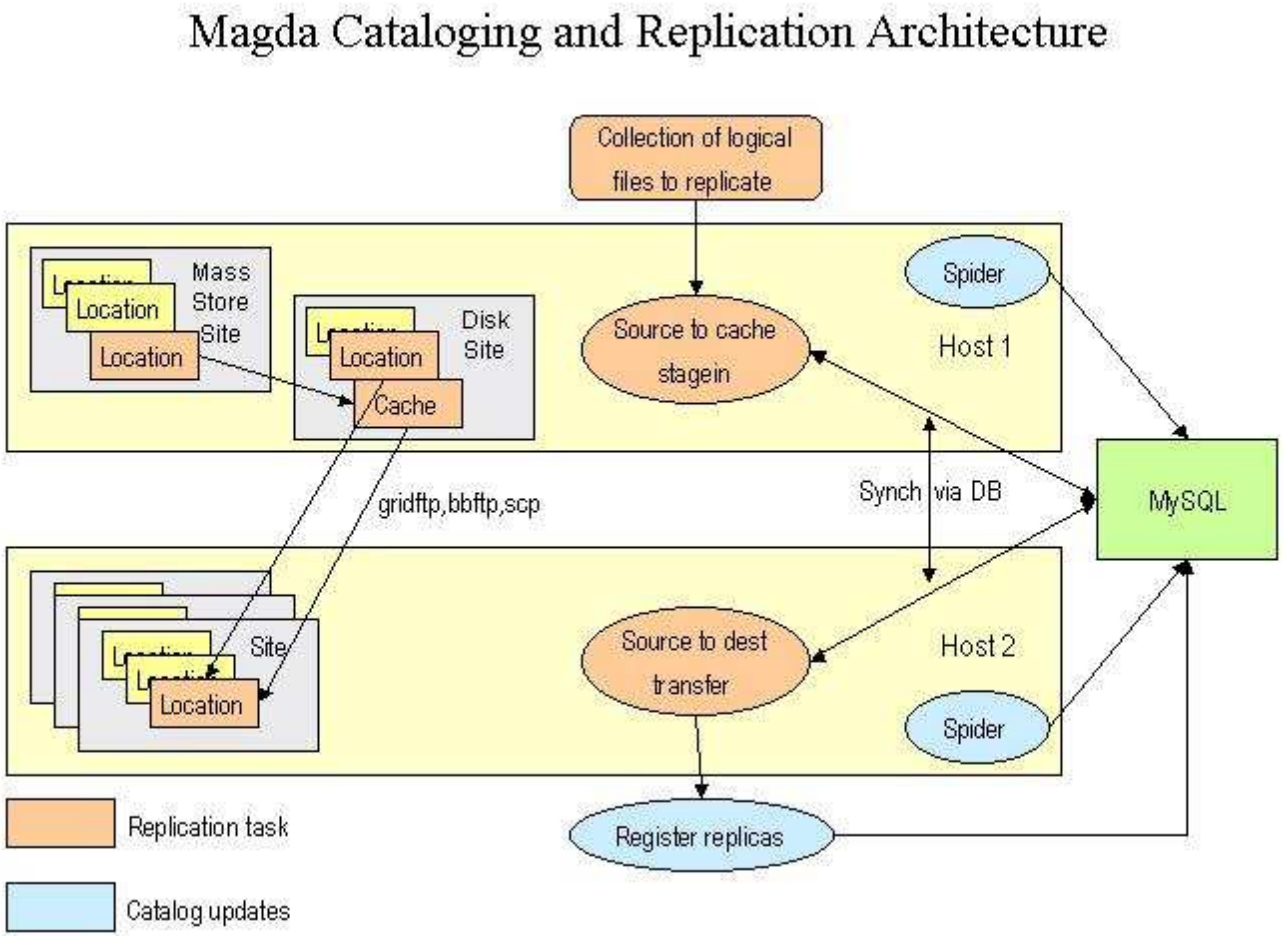
U.S. ATLAS

Capabilities of Magda

- File catalog: a file spider collects file info and fills db, files can be registered from command line also
- File movement: magda_putfile, magda_getfile, and bulk data replication. Third party transfers are supported.
- Datasets or collections of logical files
- Logical file catalog, metadata
- Query interface: web interface, magda_findfile
- Summary info: web page

U.S. ATLAS

Magda Architecture



U.S. ATLAS

What has been done in MAGDA

- Cataloging 250K file records. The file spider runs daily at CERN and US testbed sites. 23K DC1 files available through Magda.
- Has transferred ~ 3 TB data between cern castor and BNL HPSS since the start of DC1.
- Has been used in the DC1 production on grid testbed in US. Third party transfer, files put onto HPSS directly, and registered in the database.
- Command line tools available on BNL nodes for interactive users. Shared disk cache supported.

U.S. ATLAS

DIAL

- (Distributed Interactive Analysis of Large datasets)
- User
  - Works inside a C++ analysis framework
    - e.g. ROOT
  - Creates a job description
    - Specifies application (e.g. athena)
    - Creates task
      - result definition (e.g. an empty histogram, or event list)
      - code to apply to each event and fill result
    - Specifies a dataset (which event data to process)
  - Submits job description to a scheduler for interactive analysis

U.S. ATLAS

Datasets and DIAL in DC1

- Both can easily be ready for DC1 phase 2
  - prototypes for phase 1
- Datasets can provide means to
  - catalog DC1 data
    - Extension of existing “datasets”
    - Give users easy access to particular data samples
  - identify sub-samples (event selections)
- DIAL can provide
  - distributed interactive analysis of large samples
    - Event selections, filling of histograms or ntuples

U.S. ATLAS

Grappa

Grid portal for job configuration, submission, monitoring 'Athena Notebook'

- Built from Jython scripts using java-based grid tools
- Framework for web-based job submission
- Flexible: user-definable scripts saved as 'notebooks‘
- Interface/file registry to magda
- Grid monitoring
- Parameter tracking

U.S. ATLAS

Typical User session in Grappa

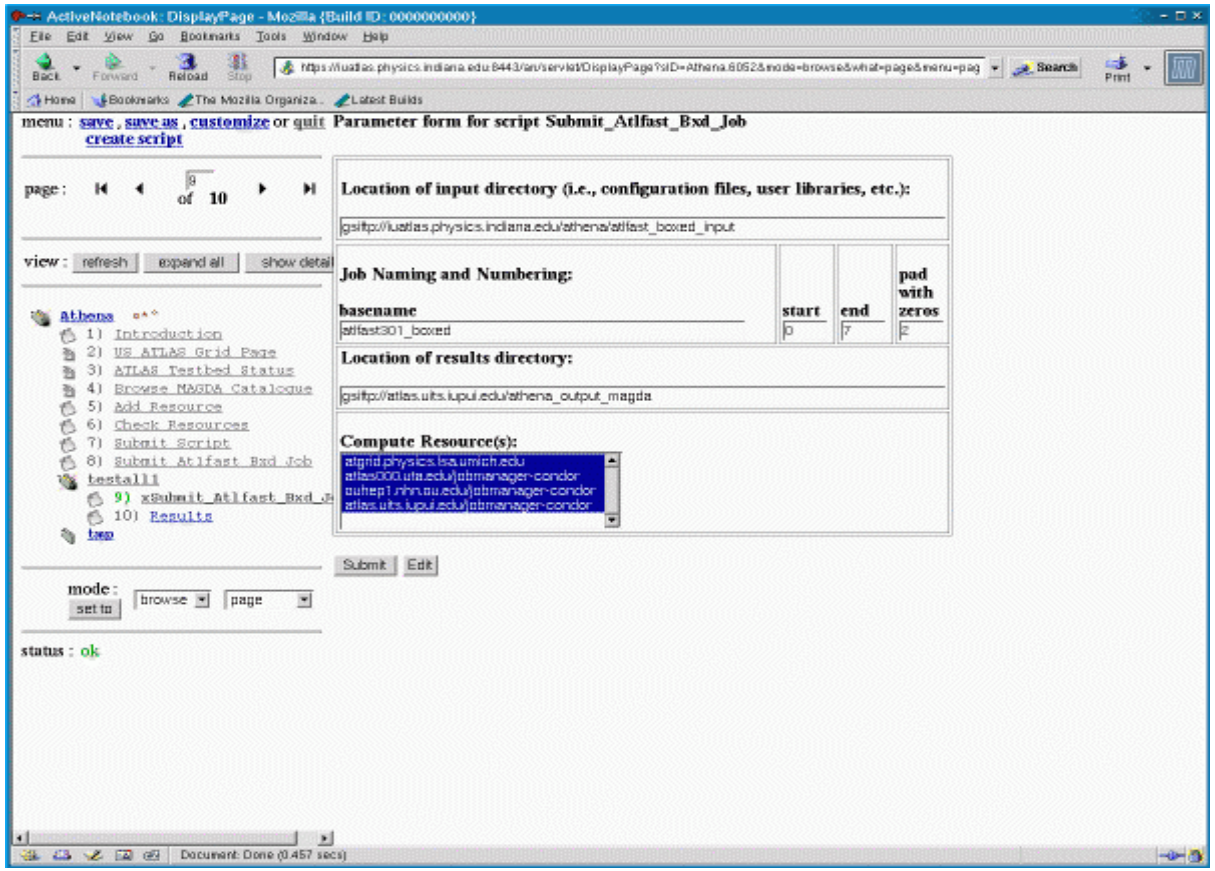
Start up portal on selected machine

Start up web browser

Configure/Select testbed resources

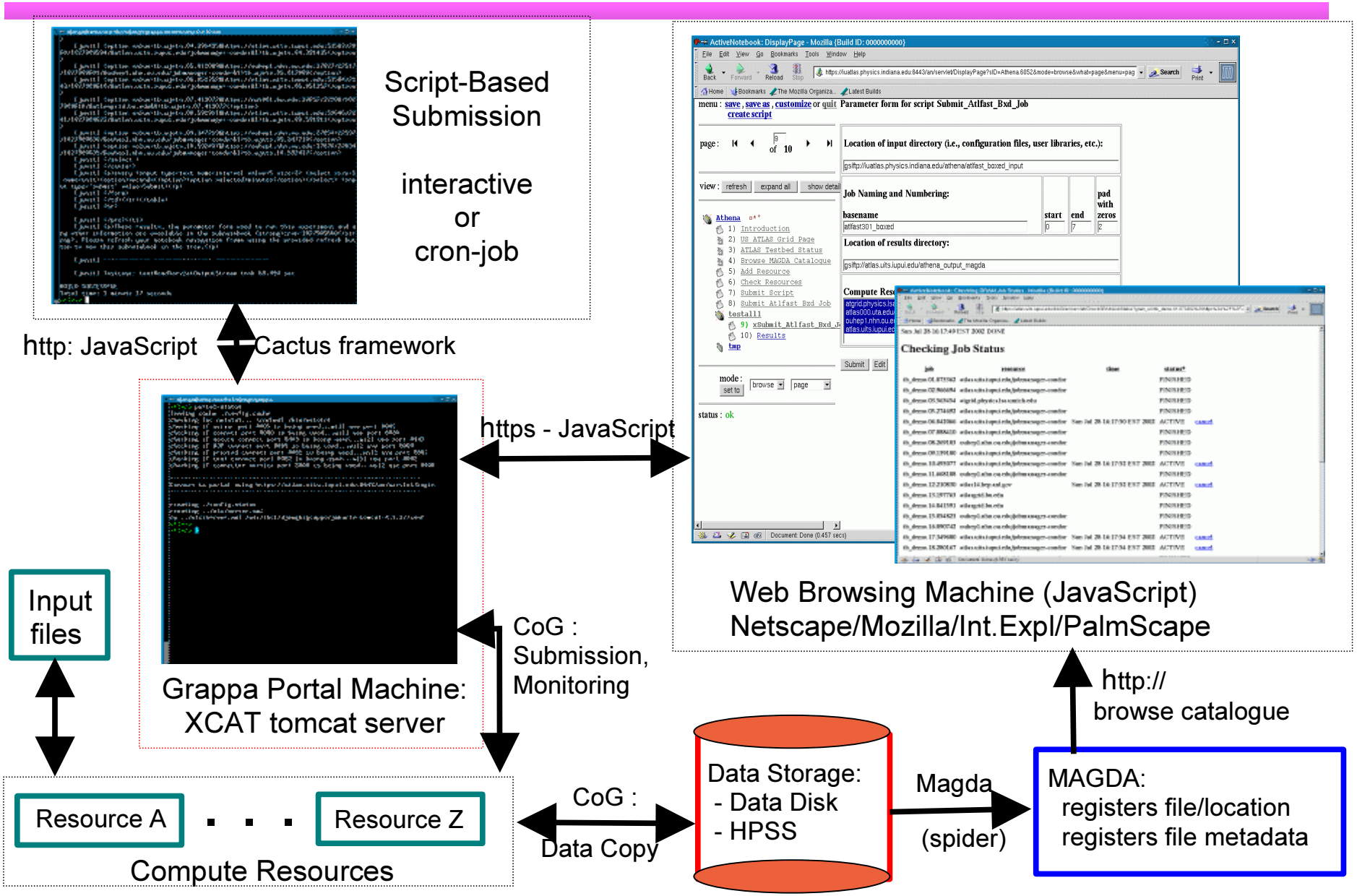
Configure input files.

Submit job



U.S. ATLAS

Grappa communications flow



U.S. ATLAS

Grappa demo

- Dan Engh (U. Chicago) will give a demonstration of grappa in the US ATLAS testbed tomorrow morning at 8:30 in the meeting hall (before the start of the regular meeting)

U.S. ATLAS

GRAT

- GRid Applications Toolkit
- Based on Globus, Magda & MySQL
- Shell & Python scripts
- Rapid development, modular design
  - parts can be run independently
- Test grid middleware, test grid performance, useful as diagnostic tool
- Includes ~50 command-line tools for
  - Athena-atlfast job submission (binary or afs)
  - Athena-atlfast production system (large scale job submission on grid, data management)
  - dc1 production and data management
  - GRAT modules will be eventually replaced by more advanced tools like Condor-G, replica catalogue, job scheduler… easy to replace

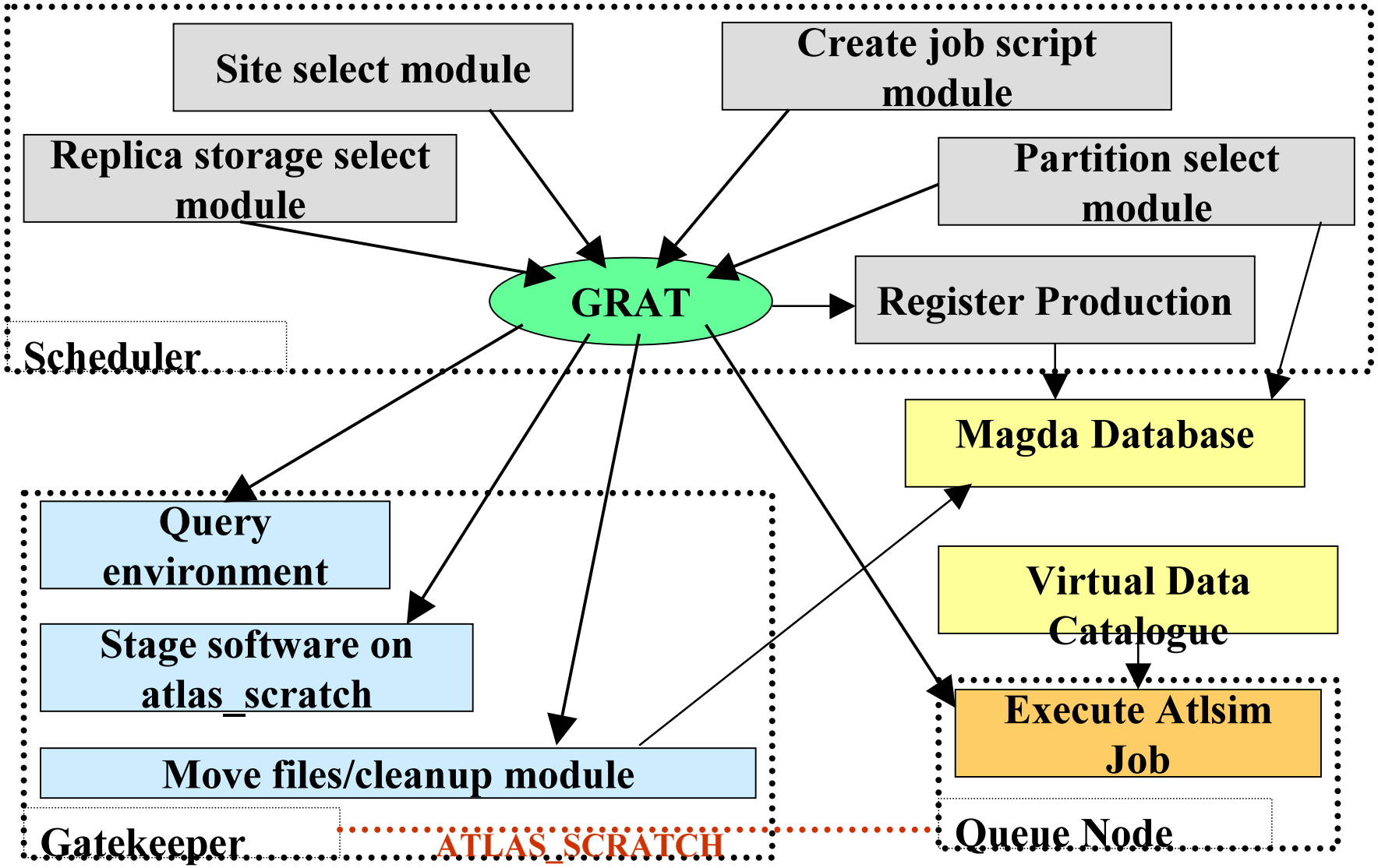
U.S. ATLAS

GRAT in DC1 Production

- **Main features:**
  - **Automatic job submission under full user control**
    - One, many or infinite sequence of jobs
    - One or many sites (for scheduling or production)
  - **Independent data management scripts to check consistency of production semi-automatically**
    - query production database
    - check Globus for job completion status
    - check Magda for output files
    - recover from many possible production failures
  - **New (and old) Magda features used:**
    - moving and registering output files stored at BNL HPSS and at replica locations on the grid
    - tracking production status

U.S. ATLAS

GRAT Job Scheduling



U.S. ATLAS

Grid Monitoring

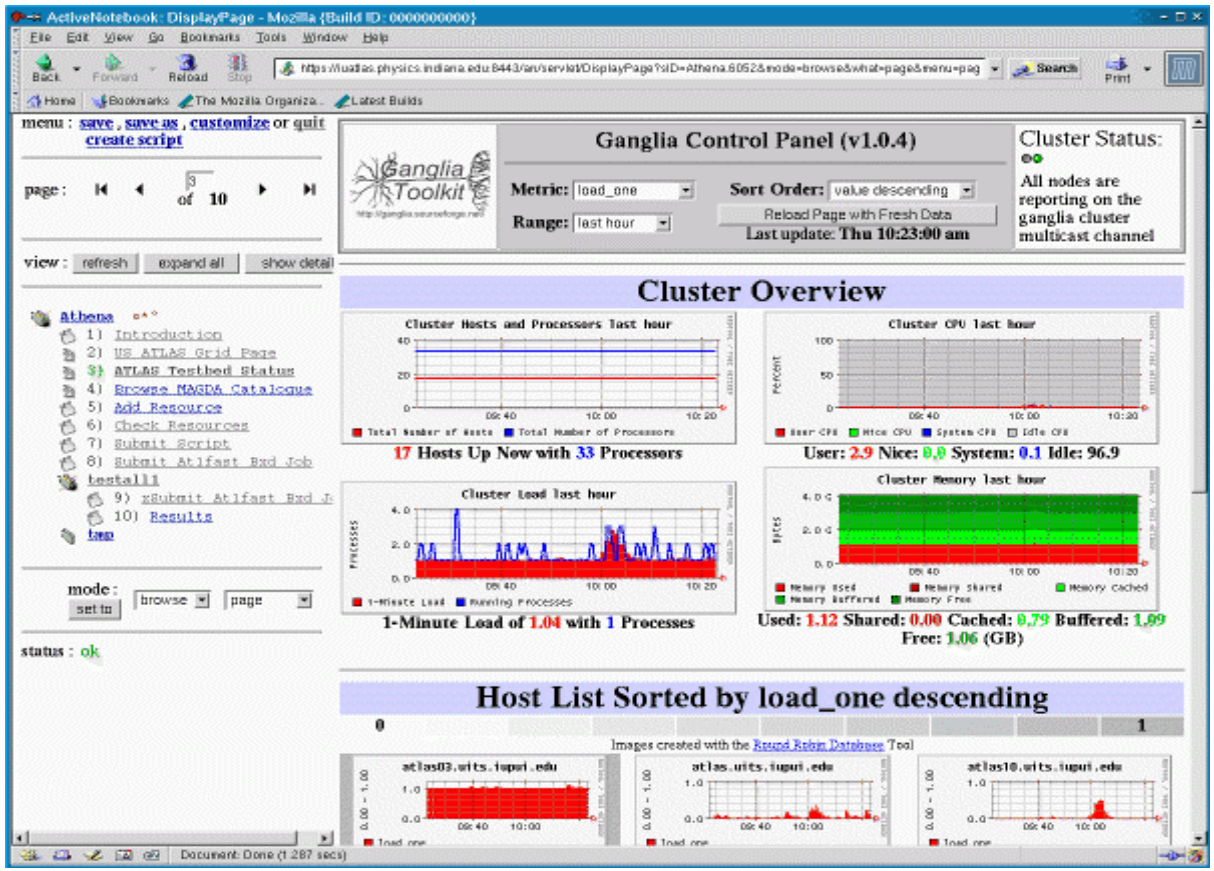
- Monitoring is critically important in distributed Grid computing
  - check system health, debug problems
  - discover resources using static data
  - job scheduling and resource allocation decisions using dynamic data from MDS and others
- Testbed monitoring priorities
  - Grid status/operations monitoring
  - Discover site configuration
  - Discover software installation
  - Application monitoring
- Also need
  - Well defined data for job scheduling
  - Visualization

U.S. ATLAS

Grid Monitoring in Grappa

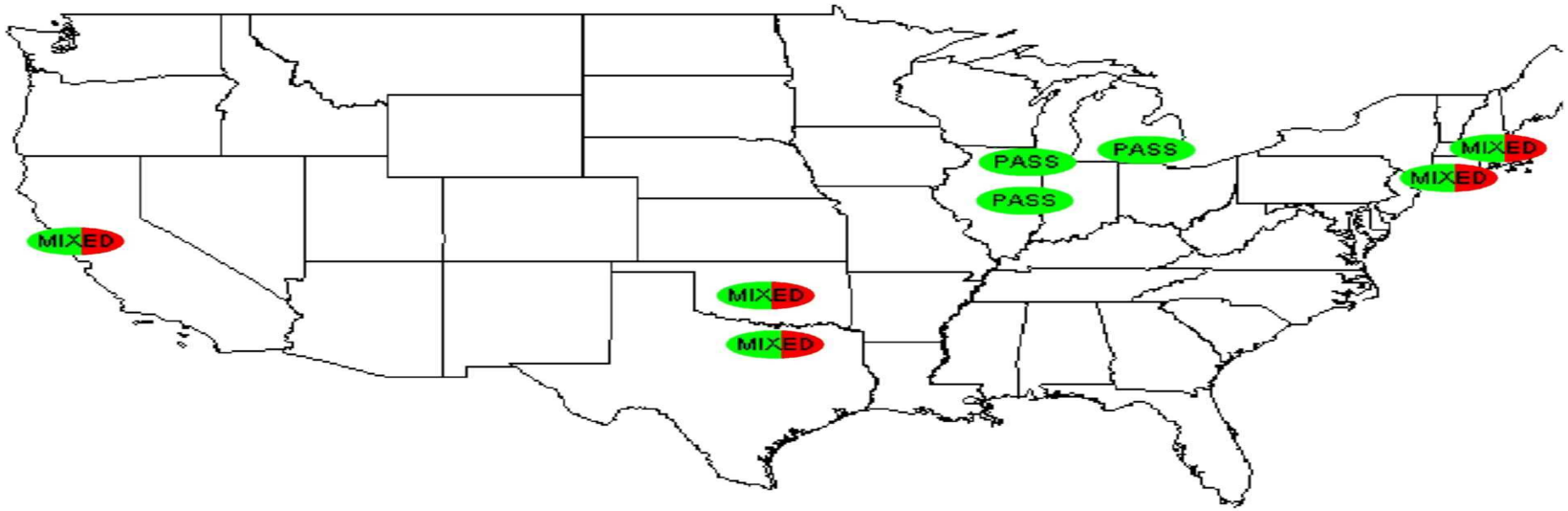
Monitor Cluster health

The new ganglia version creates an additional level combining different clusters into a “metacluster”



U.S. ATLAS

GridView



- Java applet
- Geographic view - status snapshot
  - Green - passed all grid services tests
  - Green+Red - failed some tests
- Other views
  - Hierarchical - grid services test report
  - Browser - LDAP browser for MDS information

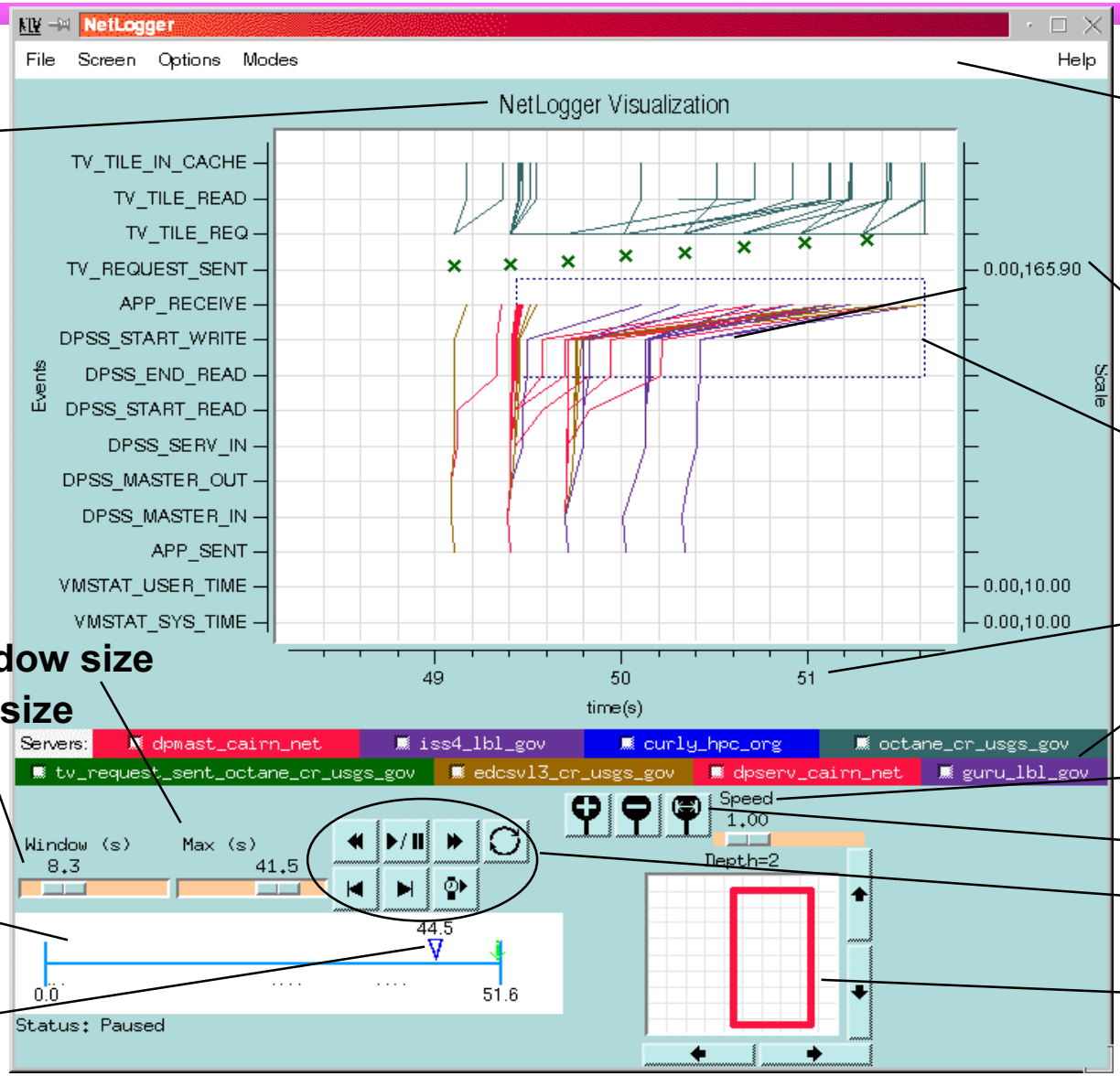
U.S. ATLAS

Athena in Netlogger

- Toolkit developed by C. Tull/B. Tierney (LBNL)
- Monitors *inside* applications run on grid
- CPU usage of subtasks monitored as “events” occur

U.S. ATLAS

NLV Analysis Tool: Plots Time vs. Event Name



Title

Events

Max window size

Window size

Summary line

You are here

Menu bar

Scale for load-line/ points

Zoom box

Time axis

Legend

Playback speed

Zoom-box actions

Playback controls

Zoom window controls

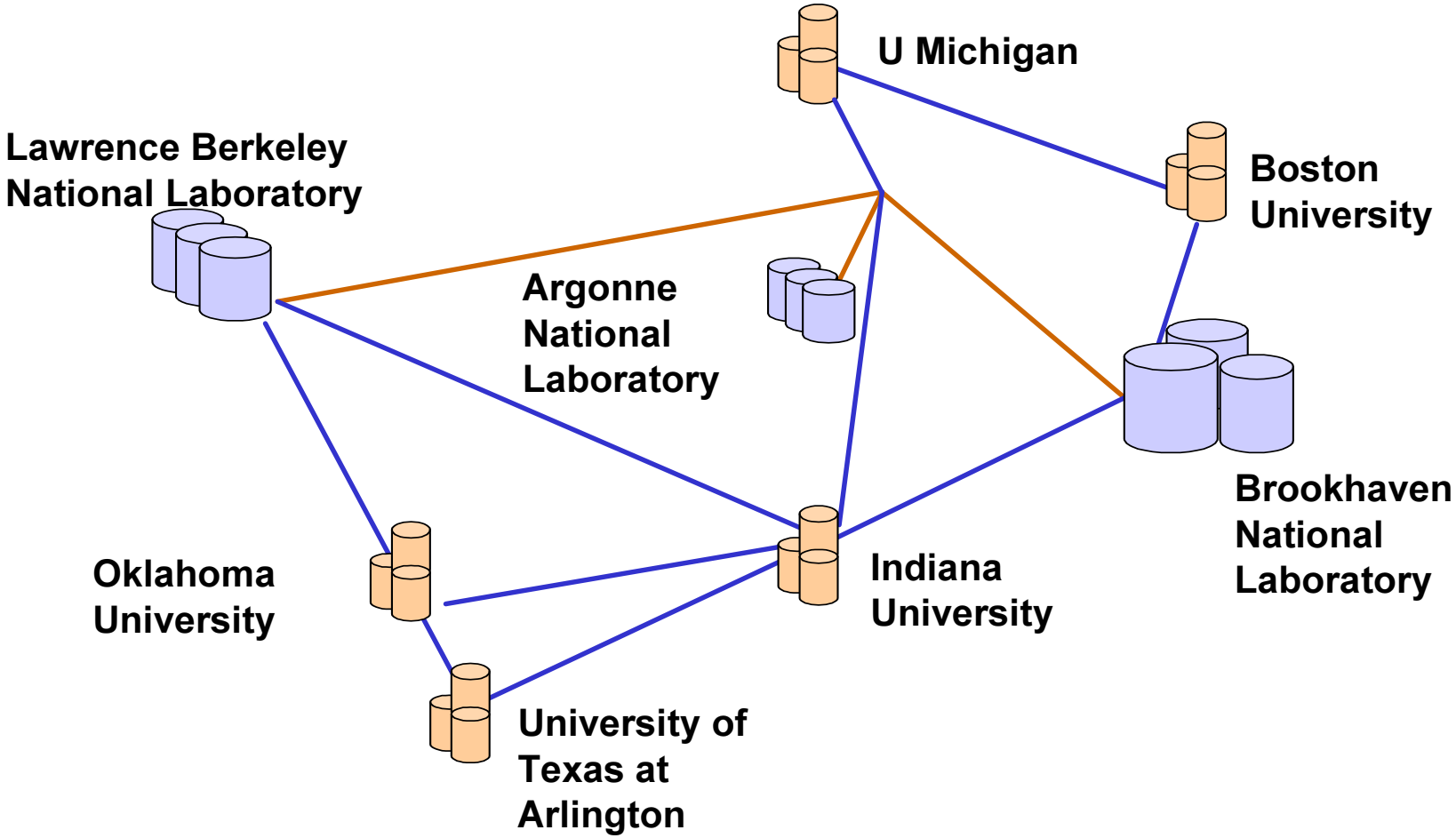
U.S. ATLAS

US Production Testing

- Summer production summary
  - June 2002 - interoperability tests
  - July 2002 - athena-atlfast production
  - August 2002 - dc1 production
- Software distribution
- GRAT/Grappa testing
- MC production experience
- dc1 phase 2 plans
- Grid monitoring tools
  - back-end information service providers
  - front-end user interface

U.S. ATLAS

Grid Testbed Sites



U.S. ATLAS

Testbed Fabric

- 8 production gatekeepers - ANL, BNL, LBNL, BU, IU, UM, OU, UTA
- Large clusters at BNL, LBNL, UTA
  - Heterogeneous system - Condor, LSF, PBS
  - LBNL: pdsf shared system, >300 nodes (average usage ~40 nodes for dc1 phase 1)
  - BNL: rcf shared system, ~100 nodes for ATLAS
  - UTA: 25 nodes dedicated for ATLAS, ~100 shared ATLAS/D0 (adding 100 more in Fall '02)
- Medium size clusters at BU, IU, UM, OU
  - 32 node systems at IU and BU, 15-20 node systems at UM and OU
  - more nodes could be added, if necessary

U.S. ATLAS

Progress since May software week

- June 2002
  - Tested interoperability - successfully ran athena-atlfast jobs on CMS & D0 & EDG grid sites
- July 2002
  - New production software released & deployed
  - 2 week Athena-atlfast MC production using GRAT & GRAPPA
  - Generated 10 million events, thousand files catalogued in Magda, all 8 sites participating
- August 2002
  - 3 week dc1 production run
  - Generated 200,000 events, using over 30,000 CPU hours, 2000 files, 100 GB storage
  - Primary files stored at BNL HPSS, replicas stored at grid testbed sites
  - Deployed VO server at BNL Tier I facility

U.S. ATLAS

DC1 Production Experience

- Anything that can go wrong, WILL go wrong
  - During 18 days of grid production (in August), every system died at least once
  - Local experts were not always be accessible (many of them on vacation)
  - Examples: scheduling machines died 5 times (thrice power failure, twice system hung)
  - Network outages - multiple times
  - Gatekeeper - died at every site at least 2-3 times
  - Three databases used - production, magda and virtual data. Each died at least once!
  - Scheduled maintenance - HPSS, Magda server, LBNL hardware, LBNL Raid array…
- These outages should be expected on the grid - software design must be robust
- We managed > 100 files/day (~80% efficiency) in spite of problems!

U.S. ATLAS

DC1 Production Statistics

$ ./dc1-list-statistics
Tue Sep 10 00:31:18 CDT 2002

total number of partitions in hpss = 1757
total size of zebra files in hpss = 42519 MB

	Files	Size
1 TeV elec. (2107)	500	13488M
1 TeV photon (2117)	500	11564M
1 TeV photon (2127)	384	8863M
1 TeV photon (2137)	371	8542M

files @ UTA	614	16100M
files @ Oklahoma	309	7191M
files @ BNL	4	113M
files @ LBNL	854	19718M

U.S. ATLAS

Lessons Learned

- Globus, Magda and Pacman make grid production easy!
- On the grid - submit anywhere, run anywhere, store data anywhere - really works!
- Error reporting, recovery & cleanup very important - will always lose/hang some jobs
- Averaged > 100 successful dc1 jobs/day with 3 grid testbed sites (other sites running atlfast production in preparation for SC2002)
- cpu usage: 8-16 hours/job of 100 events
- Found many unexpected limitations, hangs, software problems - will provide feedback to Gobus, Condor teams soon

U.S. ATLAS

Comparison of US/EDG/Nordu Grids

Topology

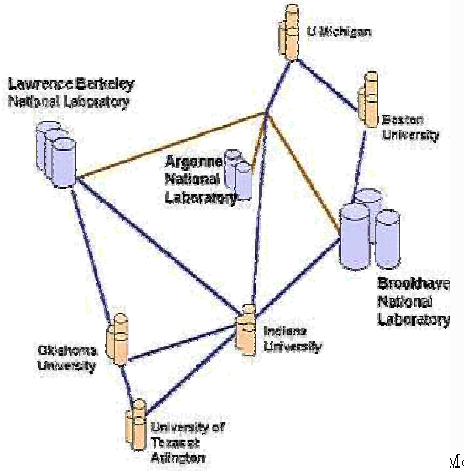
- Census of Sites
- Operating Systems
- Job Description Languages
- Job Schedulers
- Grid Middleware
- ATLAS Application Software
- User Certificates
- Signing Policies
- Brokering

U.S. ATLAS

Comparison of Topologies

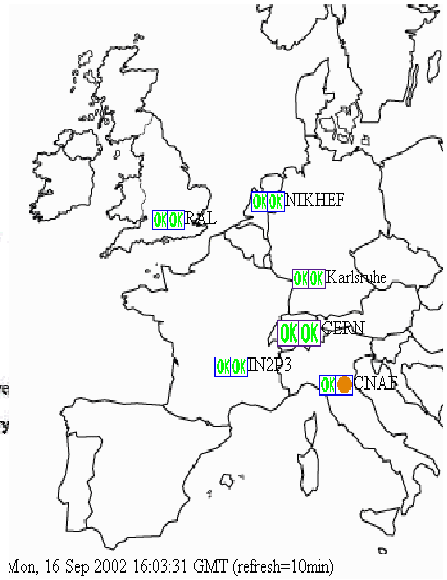
US-ATLAS

Mesh



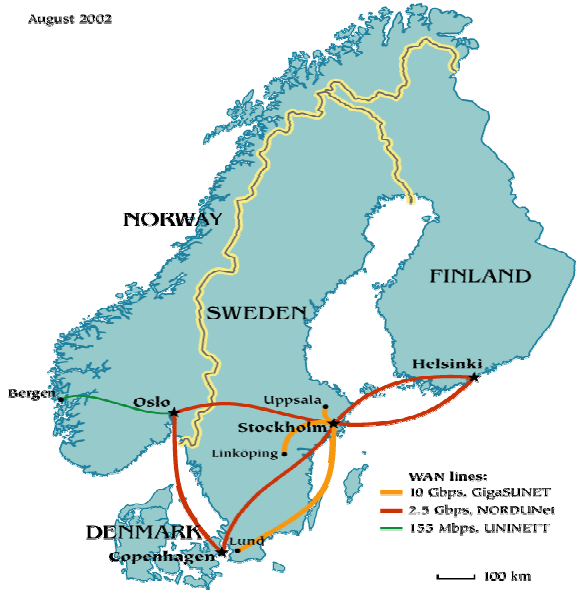
EDG Testbed Prod

Star



NorduGrid

Mesh



U.S. ATLAS

Site Census

US-ATLAS

Eight Sites: BNL, UTA, IU, BU, LBL, ANL, UofM, OU

Two nodes at ANL site run EDG Release software. One runs UI + Gatekeeper software and the other runs CE software.

EDG Testbed Prod

Six Sites: CERN, CNAF, Lyon(IN2P3), RAL, NIKHEF, Karlsruhe

NorduGrid

Twelve Sites:

Denmark – LSCF, NBI, NORDITA

Norway – Bergen Univ., Oslo Univ., Parallab

Sweden – Uppsala Univ., Lund Univ., NSC Grendel, NSC Ingvar, PDC, SCFAB

U.S. ATLAS

Job Description

US-ATLAS

All jobs are submitted using RSL (Globus Resource Specification Language) through Globus job manager, with no direct interface to queuing system

EDG Testbed Prod

All jobs are submitted using JDL (EDG Job Description Language), modified globus

NorduGrid

All jobs are submitted using XRSL (NorduGrid Extensions to the standard Globus Resource Specification Language)

U.S. ATLAS

Job Schedulers

US-ATLAS

fork, LSF, CONDOR

EDG Testbed Prod

fork, PBS, LSF

NorduGrid

mostly OpenPBS, some PBS Pro, some PBS + Maui

U.S. ATLAS

Grid Tools

US-ATLAS

- **Globus 2.0beta1 and PACMAN**
- **Software distributed by PACMAN in RPM format**
- **One Globus gatekeeper at each site**

EDG Testbed Prod

- **EDG Release 1.2.x – based on Globus 2.0 (some specific changes)**
- **Software distributed by LCFG in RPM format**
- **Globus gatekeepers only at RB, CE, and SE servers at each site**
- **Information Services and Site Statuses are available from the DataGRID WP6 site**

NorduGrid

- **NorduGrid toolkit – based on Globus 2.0 (some specific changes)**
- **Software distributed in RPM and SRPM packages**
- **Load monitor, Information system browser and Replica Catalog browser are available via the NorduGrid site**
- **One Globus gatekeeper at each site**

U.S. ATLAS

ATLAS Software

US-ATLAS

- ATLAS Release 3.2.1 installed via PACMAN – some sites can access software via AFS
- GRAT (GRid Application Toolkit) - a set of command line tools for ATLAS
- GRAPPA (GRid Access Portal for Physics Applications) - a GUI tool web portal with Athena notebook

EDG Testbed Prod

- ATLAS Release 3.2.1 installed at ATLAS-targeted Worker Node (WN) at each site
- GENIUS web portal installed at only a few sites

NorduGrid

- ATLAS 3.2.1 RPMs almost everywhere, and all sort of local application software at each site.

U.S. ATLAS

Certificates

US-ATLAS

- Globus certificates most commonly used
- DOESciencegrid certificates designated to be the standard
- Grid-mapfiles locally created and administered
- Certificates mapped to a distinct login name or group login name

EDG Testbed Prod

- Certificates issued by authorized EDG CAs accepted
- Globus certificates not accepted
- DOESciencegrid certificate accepted
- ATLAS grid-mapfile entries automatically generated from LDAP VO for ATLAS

NorduGrid

- Certificates issued by NorduGrid CA recognized at all sites
- Some sites recognize DOESciencegrid certificates – some recognize Globus
- ATLAS grid-mapfile entries automatically generated from LDAP VO for ATLAS

U.S. ATLAS

Signing Policies

US-ATLAS

- All sites recognize both the Globus and DOESciencegrid policies.
- A few sites recognize other US-only signing policies.

EDG Testbed Prod

- Each participating EDG member has their own policy.
- All sites recognize the policies from all participating EDG members.
- All sites recognize the DOESciencegrid policy.
- The Globus policy is not recognized at any site.

NorduGrid

- Each site has it's own policy.
- The NorduGrid CA policy is recognized by all.
- EDG Testbed member CA policies are recognized by some.
- Some sites recognize the Globus policy.

U.S. ATLAS

Brokering

US-ATLAS

- No automated resource brokering
- Information servers starting to be deployed
- Currently, job submittals rely on user knowledge of resources

EDG Testbed Prod

- Two Resource Brokers for Testbed – provides for fault tolerant backup of brokering
- Monitor and Discovery Services (MDS) provided by servers in each participating country

NorduGrid

- Brokering is performed by User Interfaces using input from the NorduGrid Information System service.

U.S. ATLAS

Data Management

US-ATLAS

- File Replication and Management - Magda
- Catalog Services – MySQL
- User Interfaces – web-interface/command line/C++, Java, Perl APIs

Merging GDMP and Globus Replica Catalog functionality into Magda is actively being worked.

EDG Testbed Prod

- File Replication and Management - GDMP 3.0
- Transition to Reptor
- Catalog Services – Globus Replica Catalog
- User Interfaces – command line/C++ APIs

NorduGrid

- File Replication and Management - GDMP 3.0 + local patches
- Catalog Services – Globus Replica Catalog
- User Interfaces – command line/C++ APIs

U.S. ATLAS

DC1P2 Comments

- Is it possible to have a completely uniform set of grid tools for DC1P2? Differences:
  - Globus differences US/EDG/Nordu
  - Job schedulers
  - Job submissions
  - Certificate authority
  - Data management
- Minimal requirements for grid production
  - Uniform ATLAS release – no “site tuning”
  - Uniform tools for file registration
  - Pacman+Magda as tools

U.S. ATLAS

ATLAS SW Distribution for DC1 phase 2

- **The ONE Official build, made by the Librarian, must be the one that is distributed for GRID production.**
  - **The 3 commands:**
    1. **cmt broadcast cmt config**
    2. **source setup.csh**
    3. **cmt broadcast gmake**
    - **MUST make the official release, e.g. rpm files**
    - **This requires some support from the cmt team**
    - **There must be an exportable run time environment**
  - **ONE command must install the sw at a remote site**
    - **e.g.**
      - **Pacman –get ATLAS_DC1p2**
        - **Once rpm files are made by cmt, this is a trivial matter of writing a small pacman cache file**

U.S. ATLAS

DC1 Phase 2 in US

- **Pile up production:**
  - plan to start mid-october
  - grid-based production using 200-300 nodes
- **Athena reconstruction**
  - grid testbed will participate
  - store ESD/AOD at BNL, Ntuples on grid testbed
  - need grid deployable athena package!
- **Analysis / user access to data**
  - magda_getfile already provides access to ~30k catalogued dc1 files from/to many grid locations (need ATLAS VO to make this universal)
  - long term: DIAL being developed at BNL http://www.usatlas.bnl.gov/~dladams/dial/

U.S. ATLAS

Long Term: complementary efforts

- EDG strengths
  - Resource Brokering
  - VO software
  - Use of MDS in batch jobs
  - Portal submissions
- US strengths
  - Package distribution
  - File registration
  - Virtual data
  - Portal submissions

U.S. ATLAS

A Path

- ATLAS and LCG need to fashion a coherent list of grid tools
  - Can gain from the experience of three testbeds
  - Further divergence will make life difficult (maybe we have to live with it)
- Partition effort among developers to avoid duplication
- Coherence with application developers important
  - But factorization of problems is the key