# U.S. ATLAS Grid Testbed Status

**Kaushik De**

**University of Texas at Arlington**

**PPDG Meeting**

**BNL, September 2002**

# Outline

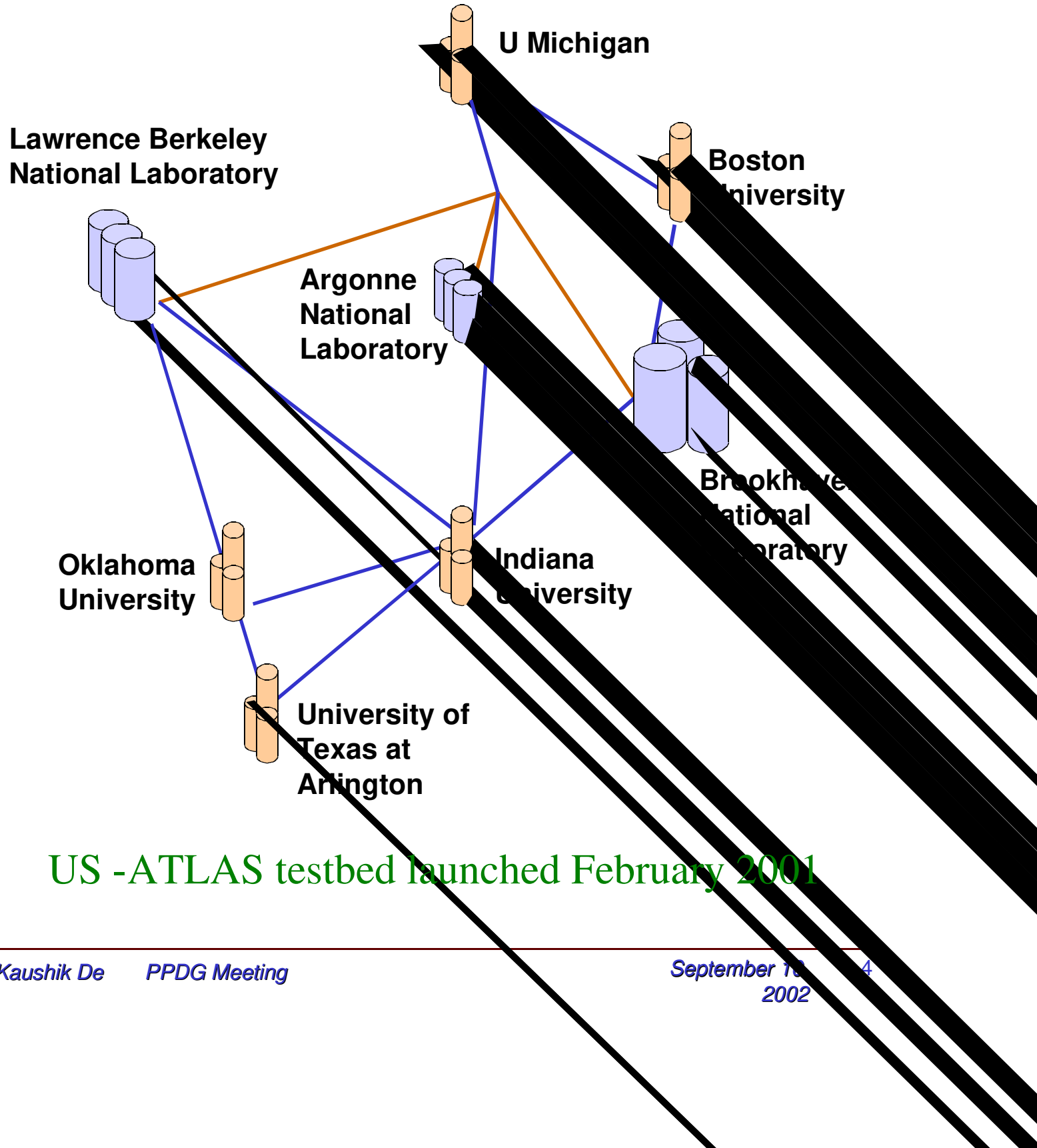
- Testbed introduction
- Recent accomplishments
- Software distribution
- Application toolkit
- MC production
- Monitoring
- Grid tools
- Integration

✳ During summer 2002, U.S. Grid Testbed focussed on rapid software deployment and distributed Monte Carlo production

# Testbed Goals

- Demonstrate success of grid computing model for High Energy Physics
  - in data production
  - in data access
  - in data analysis
- Develop, deploy and test grid middleware and applications
  - integrate middleware with applications
  - simplify deployment - robust, rapid & scalable
  - inter-operate with other testbeds & grid organizations (iVDGL, DataTag…)
  - provide single point-of-service for grid users
- Evolve into fully functioning scalable distributed tiered grid

# Grid Testbed Sites




US -ATLAS testbed launched February 2001

# Testbed Fabric

- 8 production gatekeepers - ANL, BNL, LBNL, BU, IU, UM, OU, UTA
  - http://heppc1.uta.edu/atlas/grid-testbed/testbed-sites.htm
- Large clusters at BNL, LBNL, IU, UTA, BU
  - Heterogeneous system - Condor, LSF, PBS
  - Currently > 100 nodes available
  - Could double capacity quickly, if needed
- + Multiple R&D gatekeepers
  - gremlin@bnl - iVDGL GIIS
  - heppc5@uta - ATLAS hierarchical GIIS
  - atlas10/14@anl - EDG testing
  - heppc6@uta+gremlin@bnl - glue schema
  - heppc17/19@uta - GRAT development
  - few sites - Grappa portal
  - bnl - VO server
  - few sites - iVDGL testbed

# Summer Accomplishments

## May 2002

- Globus 2.0 beta RPM developed at BNL
- Athena-atlfast grid package developed at UTA
- Software installation - Pacman developed at BU
- GRAT toolkit for job submission on grid, developed at UTA & OU

## June 2002

- Increased participation by all sites
- Tested interoperability - successfully ran ATLAS MC jobs on CMS & D0 grid sites
- ANL demonstrated that U.S. testbed package can run successfully at EDG sites

# Testbed Summer contd.

- **July 2002**
  - New production software released & deployed
  - 2 week Athena-atlfast MC production run using GRAT & GRAPPA
  - Generated 10 million events, thousand files catalogued in Magda, all sites participating

- **August 2002**
  - New production software package
  - 3 week dc1 production run using GRAT
  - Generated 200,000 events, using over 25,000 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

# Software Distribution

- Goals:
  - Easy installation
  - Uniform software versions
  - Pacman perfect for this task
- Deployment in 3 steps
  - VDT Package (root)
    - Pacman, Globus 2.0b, Condor, Pippy...
  - ATLAS Package (non-root)
    - Magda, Grappa, JDK, MySQL, Perl…
  - User Package
    - GRAT, Grappa
- Experience
  - Repeated process 3 times during summer 2002
  - Worked well!
  - Deployed successfully at all testbed sites

# Installation Web Page

## Setting up a U.S. ATLAS Testbed Site

These steps are now certified and frozen for Testbed 2 (July, 2002 production run)

*The following steps should be followed by the Grid Administrator:*

1. Saul's Installation Instructions
   - This will Install/Update Globus and Atlfast-native
   - Pippy 0.2 will be installed and configured
   - And install Atlas specific packages - Perl 5.6.0, MySQL Client 3.23.49, JDK 1.3.1_03, Magda 1.0, and Pacman 2.066
2. Check your installation
3. Compare your installation to other testbed sites

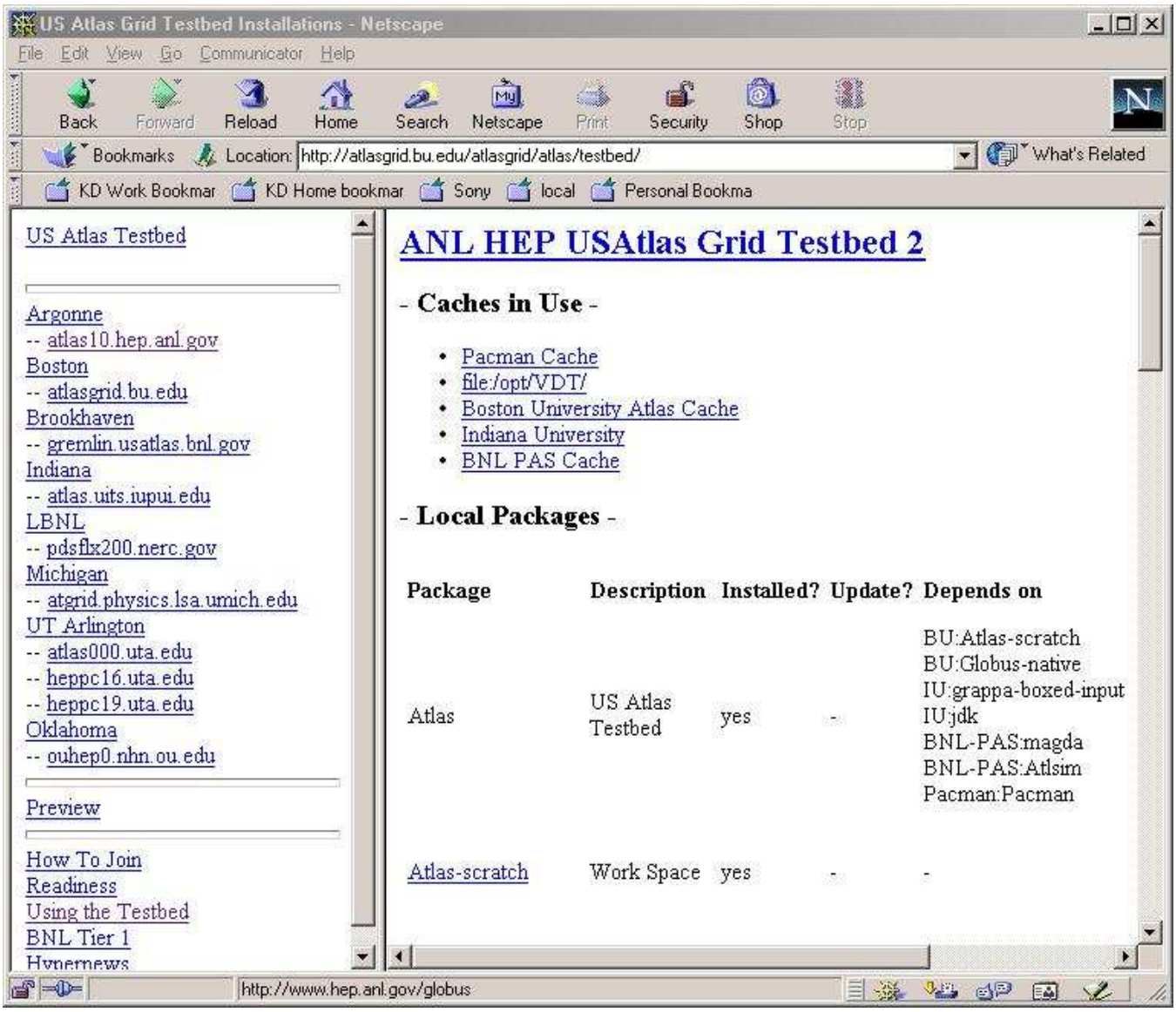
*Useful links:*

- Wensheng's Magda notes
- Instructions which were used to setup Testbed 1 in April 2002

---

*by* Kaushik De, email: kaushik@uta.edu.
Last modified on: Thursday, July 11, 2002 18:14:22. •

# Installation Status




- Installation information also available through Pippy using Globus MDS

# User Toolkits

## U.S. Atlas Grid Testbed User Software

**This software has not yet been certified for July production!**

### Installing User Tools

- Log in to your *personal user account* on a Grid Testbed Machine which is properly configured with the Atlas environment. If you are not sure, check with your system administrator. Here is a list of correctly configured machines
- Create a new working directory, for example ~atlastoys
- `$ cd` into working directory
- Set up system level Atlas software environment:
  1. For bash shell: `$ . /usr/local/bin/atlas_setup.sh`
  2. For csh/tcsh shells: `% source /usr/local/bin/atlas_setup.csh`
- `$ pacman -cache:$ATLAS_LOCATION -get BU:Atlas-user`
- Log out of your account

### Using Testbed Tools

- Two packages were installed by the above procedure - GRAT and GRAPPA
- Log in to your personal account
- cd into the working directory you created above during installation
- Set up your local Atlas software environment:
  1. For bash shell: `$ . ./setup.sh`
     (Ignore the error message setenv not found)
  2. Or for csh/tcsh shells: `% source setup.csh`
- Start application:
  1. Follow instructions at GRAT (GRid Application Toolkit) - a set of commandline tools
  2. `portal-start` will start GRAPPA (GRid Access Portal for Physics Applications) - a GUI tool

---

*by* Kaushik De, email: kaushik@uta.edu, homepage: http://heppc1.uta.edu/kaushik.
Last modified on: Wednesday, July 24, 2002 14:48:22 . •

# GRAT

- Grid Applications Toolkit
- Based on Globus, Magda & MySQL
- Shell & Python scripts
- Modular design
  - parts can be run independently
- Rapid development platform
- 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
  - dc1 production and data management
  - Modules will be replaced by more advanced applications like Condor-G, replica catalogue...

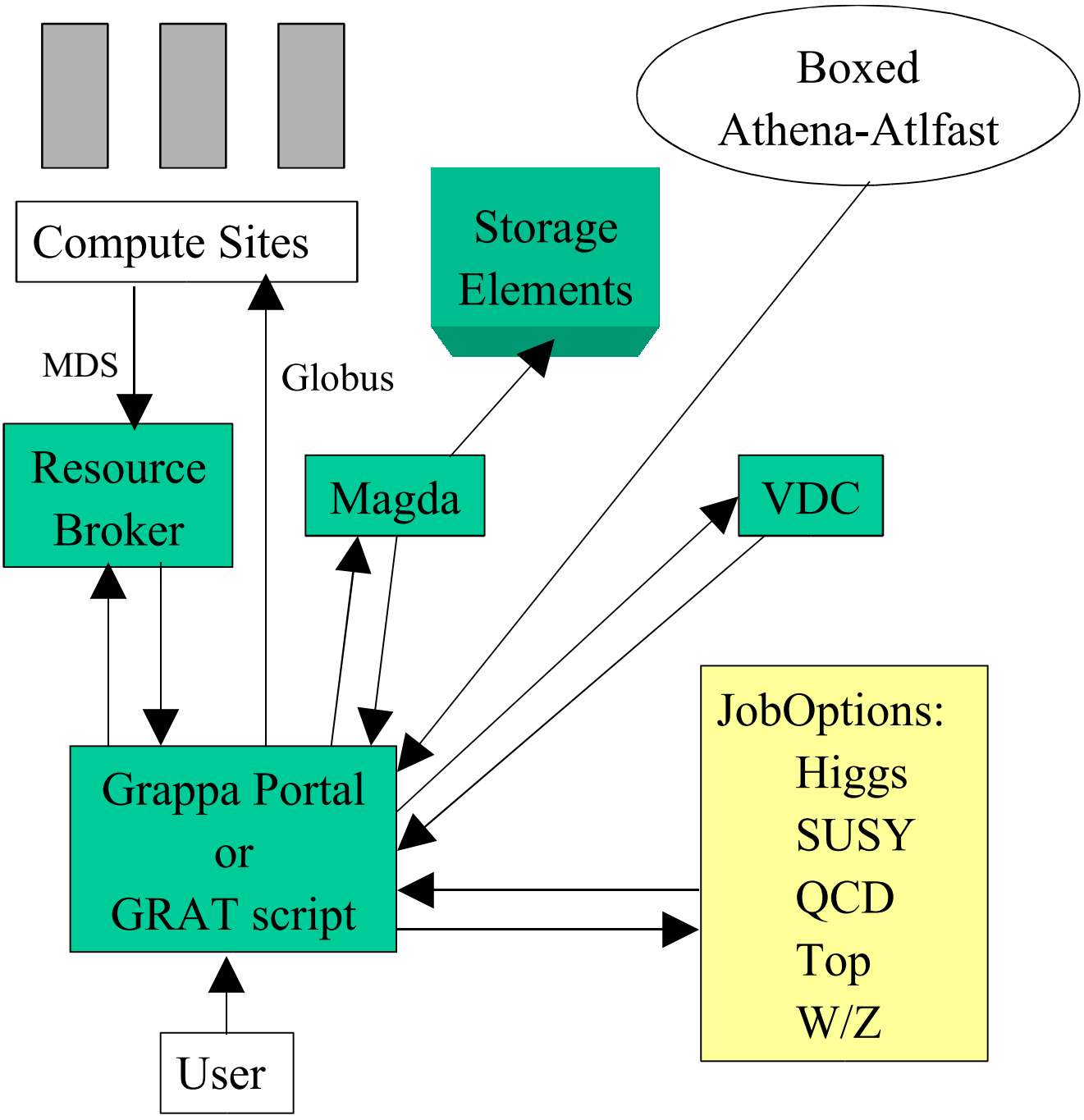
# GRAT Scripts!

atlfast-afs*
atlfast-afs-cleanup*
atlfast-afs-condor*
atlfast-afs-multi*
atlfast-afs-setup*
dc1-data-checkpart*
dc1-data-ispart*
dc1-data-missing*
dc1-data-putprime*
dc1-data-reclean*
dc1-data-verify*
dc1-data-xcheck*
dc1-data-xcheck-python*
dc1-forever*
dc1-list-currentjobs*
dc1-list-jobs*
dc1-list-localjobs*
dc1-list-statistics*
dc1-list-testfiles*
dc1-run-buffer*
dc1-run-getpart*
dc1-run-highest*
dc1-run-job*
dc1-run-logcheck*
dc1-run-movefiles*
dc1-run-notpart*
dc1-run-pending*
dc1-run-script*
dc1-submit*
gridstat*
production-checknodes*
production-cleanup*
production-execute*
production-filename*
production_gatetest.py
production-getvars*
production-jobopts*
production-listqueues*
production-makeremote*
production-primestore*
production-queues*
production-replicastore*
production-scheduler*
production-stats*
production-statsfile*
production_tcptest.py
README
run-atlfast*
run-atlfast-afs*
run-atlfast-afs-condor*
run-atlfast-afs-multi*
run-atlfast-boxed*
run-atlfast-boxed7*
run-atlfast-production*

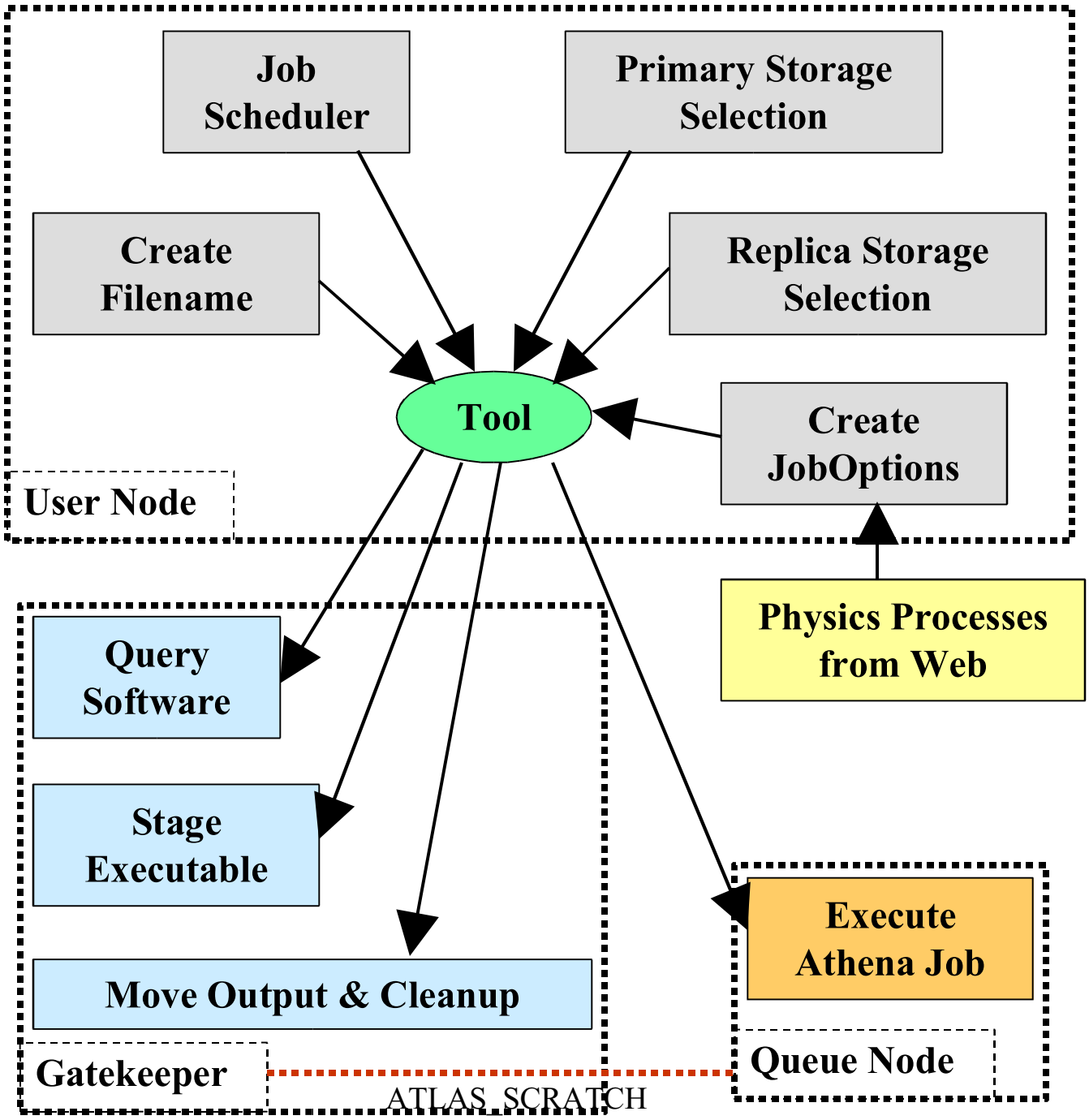
# DC1 Production Scripts

- Main tools:
  - dc1-forever: simple opportunistic scheduler
  - dc1-submit: submit multiple production jobs
  - dc1-run-*: modules used by dc1-submit
  - dc1-list-*: generate status and statistics
  - dc1-data-*: tools to manage data
- Main features:
  - Uses MySQL production database
  - Magda for data catalogue, BNL HPSS storage
  - Globus for job submission
  - gsiftp for file movement
  - multiple scheduler -> multiple gatekeepers
- Limitations:
  - Single scheduler < 128 jobs
  - Single gatekeeper < 64 jobs
  - major shutdowns required user intervention

# Atlfast Production Architecture (May '02)



# Implementation (July '02)



# Atlfast Tools

- Production toolkit examples:
  - run-atlfast-production
    - at [sites/all, # events, physics/all]
  - production-filename
    - generate base+date+events+physics+site
  - production-getvars
    - query gatekeeper for env variables
  - production-queues
    - return list of queues (static)
  - production-listqueues
    - MDS queried list of queues
  - production-primestore
    - primary magda cache location
  - production-replicastore
    - return replica location
  - checknodes
    - return available nodes (globus)
  - stats/stats2
    - magda file statistics

# Testbed Production

- Goals:
  - Demonstrate distributed ATLAS data production, access and analysis using grid middleware and tools developed by the testbed group
- Plans:
  - **Atlfast production** to test middleware and tools, and produce physics data for summer students, based on athena-atlfast fast MC
    - 2 weeks to regenerate data, once a month
    - deploy new tools and middleware each cycle
    - move away from farm paradigm to grid model
    - very aggressive schedule - people limited!
  - **DC1 production** to test fabric capabilities and produce and access data, using old Fortran code atlsim, atrig and atrecon
    - not repeatable - but real test of grid software
    - increase U.S. participation - involve grid testbed
    - CPU intensive - ~14 hours per job/file

# 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 GB

| | 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 |

# 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!

# Production Status Checking

[kaushik@atlas000 usertools-0.7]$ ./dc1-list-jobs
Checking the status of 123 globus jobs.
.....

| | Active | Pending |
|---|---|---|
| Jobs at LBNL: | 39 | 0 |
| Jobs at OU : | 8 | 0 |
| Jobs at UTA : | 24 | 6 |

[kaushik@atlas000 usertools-0.7]$ condor_q
-- Submitter: atlas000.uta.edu :

| ID | OWNER | ITTED | RUN_TIME | ST | PRI | SIZE | CMD |
|---|---|---|---|---|---|---|---|
| 2107.0 | kaushik | 9/8 12:41 | 1+02:17:43 | R | 0 | 0.0 | dc1-run |
| 2108.0 | kaushik | 9/8 12:43 | 1+02:13:29 | R | 0 | 0.0 | dc1-run |
| 2109.0 | kaushik | 9/8 12:45 | 1+02:13:27 | R | 0 | 0.0 | dc1-run |
| 2110.0 | kaushik | 9/8 13:12 | 1+01:43:58 | R | 0 | 0.0 | dc1-run |
| 2114.0 | kaushik | 9/8 13:21 | 1+01:37:06 | R | 0 | 0.0 | dc1-run |
| 2115.0 | kaushik | 9/8 13:50 | 1+01:08:37 | R | 0 | 0.0 | dc1-run |
| 2145.0 | kaushik | 9/9 14:44 | 0+00:00:00 | I | 0 | 0.0 | dc1-run |
| 2146.0 | kaushik | 9/9 14:46 | 0+00:00:00 | I | 0 | 0.0 | dc1-run |

30 jobs; 6 idle, 24 running, 0 held

# Atlfast Production

- **July 1-15: Phase 0, 10^7 events**
  - Globus 2.0 beta
  - GRAT and Grappa - using athena-atlfast 3.0.1, common disk model
  - Magda - data management, 1000 files
  - 5 physics processes - http cached
    - Use Pythia, standard ATLAS physics group jobOptions
    - 000 - atlfast default
    - 001 - minbias
    - 002 - dijets 3 GeV < $P_t$< 100 GeV
    - 003 - all jets
    - 004 - W + jets
    - 005 - Z (ll) + jets
    - 006-010 processes repeat

# Production Status

```
US_atlfast_production Summary as of Wed Jul 31 13:01:39 2002:

Minbias         : 126 files,   889700 evts.
Dijet           : 105 files,   845000 evts.
All jets        : 104 files,   851300 evts.
W + jets        : 104 files,   860100 evts.
Z to ll + jets  : 105 files,   830000 evts.


Site Breakdown:

atlas.uits.iupui.edu     :  73 files,   590100 evts.
atlas000.uta.edu         : 174 files,  1246800 evts.
gremlin.usatlas.bnl.gov  :  20 files,    19400 evts.
ouhep0                   :   2 files,      600 evts.
ouhep0.nhn.ou.edu        : 174 files,  1481200 evts.
ouhep1.nhn.ou.edu        :   1 files,     1000 evts.
pdsfgrid.nersc.gov       :  99 files,   936000 evts.
pdsflx200.nersc.gov      :   1 files,     1000 evts.


Total: 544 files,  4276100 evts.

Magda query took 10 seconds.
```

# 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
- Found many unexpected limitations, hangs, software problems - will provide feedback to Gobus, Condor, GRAT teams soon

# 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 other monitors
- Testbed monitoring priorities
  - Discover site configuration
  - Discover software installation
  - Application monitoring
  - Grid status/operations monitoring
- Also need
  - Well defined data for job scheduling
  - Visualization

# Monitoring - Back End

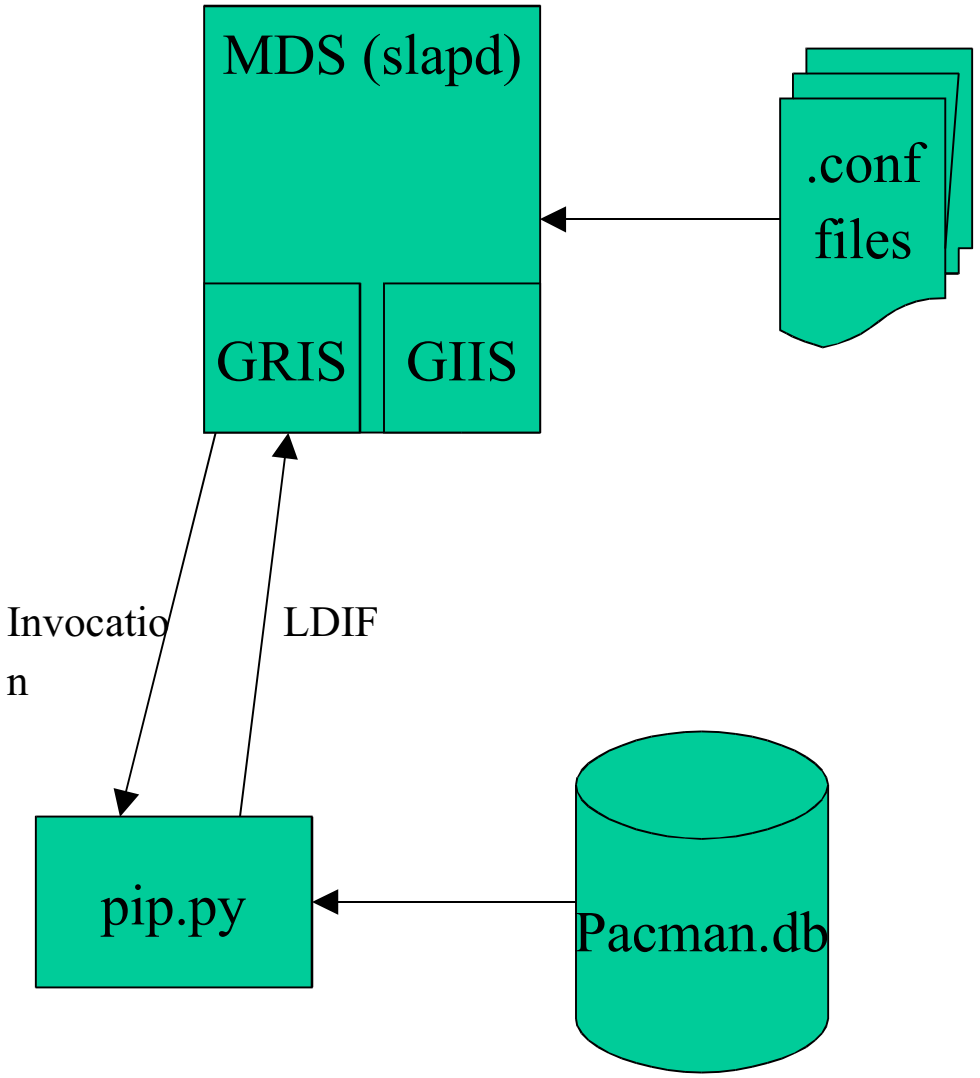
- Publishing MDS information
  - Glue schema - BNL & UTA
  - Pippy - Pacman information service provider
  - BNL ACAS schema
  - Hierarchical GIIS server
- Non-MDS back ends
  - iPerf, Netlogger, Prophesy, Ganglia
- Archiving
  - MySQL
    - GridView, BNL ACAS
  - RRD
    - Network
- Work needed
  - What to store?
  - Replication of archived information
- Good progress on back end!

# PIPPY

- Publishes information into MDS about software packages installed via Pacman.
- Uses LDAP schema elements that describe a package.
- Interfaces with GRIS backend of MDS service.

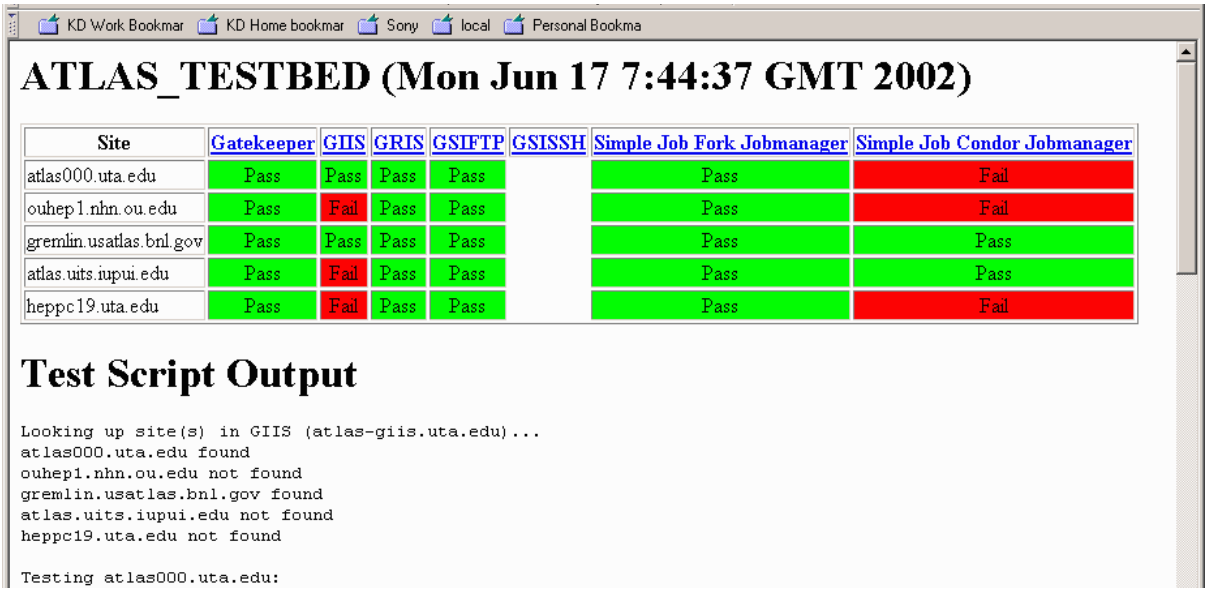
- pippy creates two type of entries:
  - Software Collection
  - Software Package
- Collection allows aggregation of packages.
- Package describes contents of a pacman file.

# Architecture

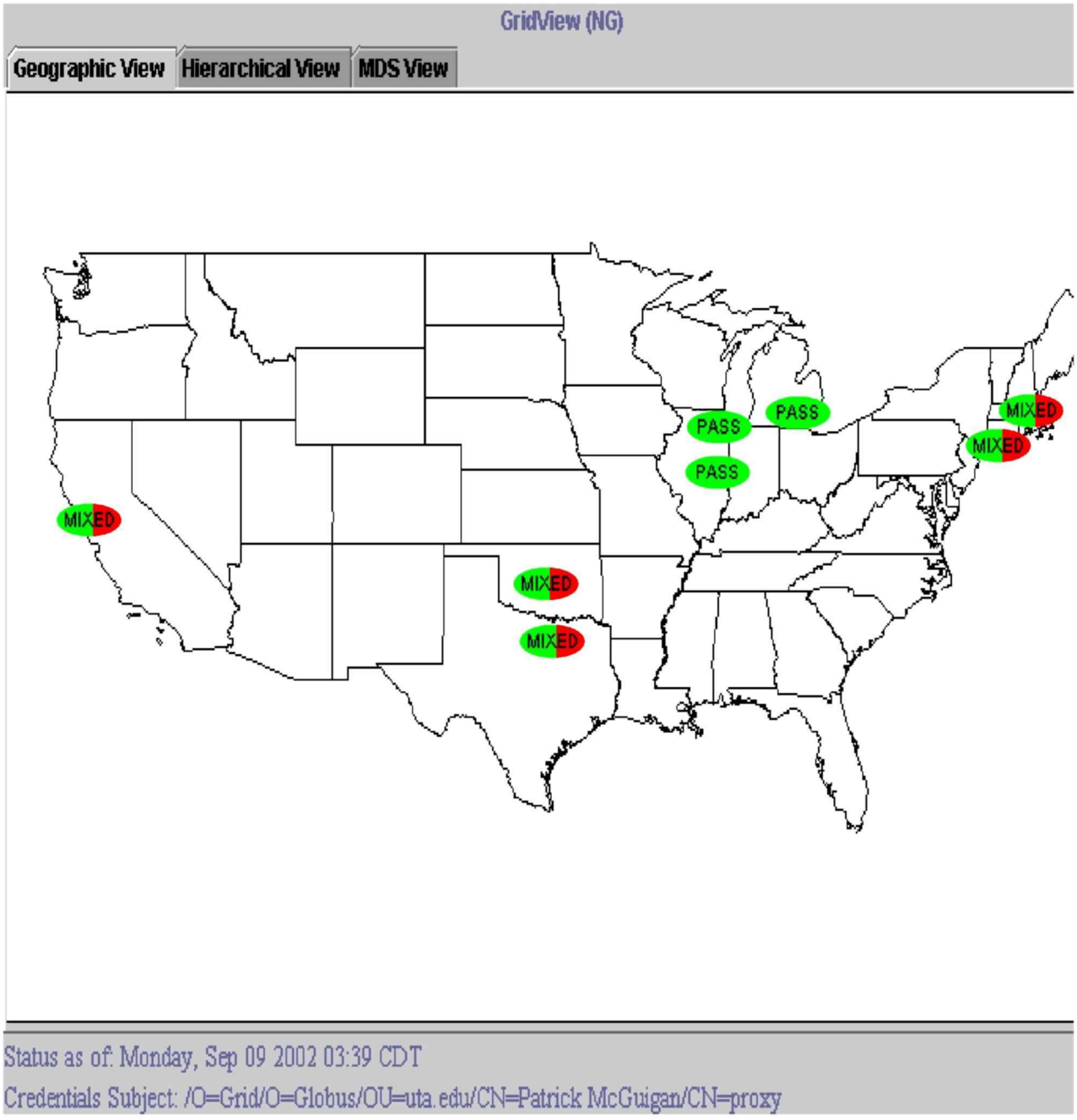


# Monitoring - Front End

- MDS based
  - GridView, Gridsearcher, Gstat
  - Converting TeraGrid and other toolkits
- Non-MDS
  - Cricket, Ganglia
- Example:
  - Gstat - text version

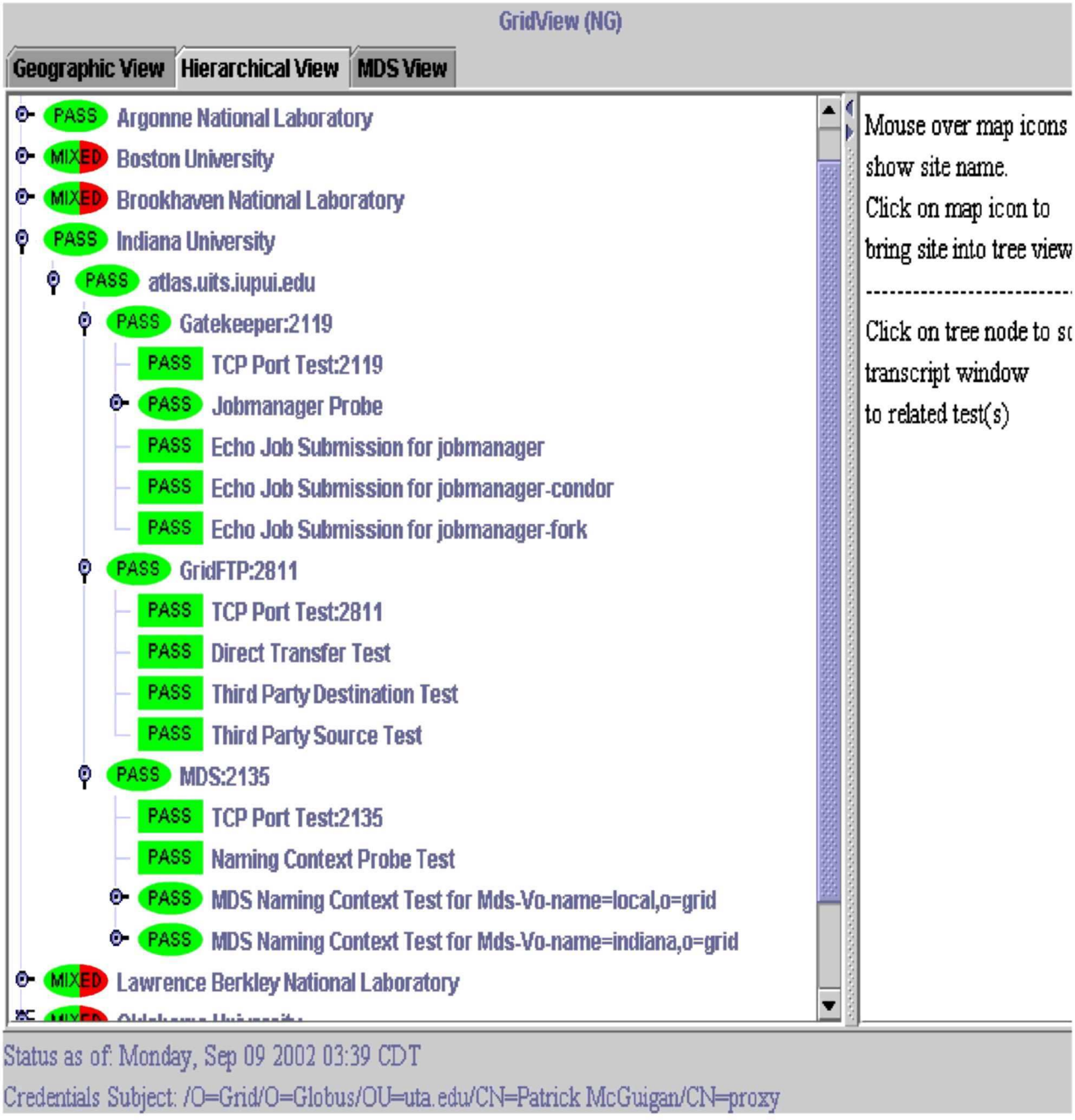


# GridView (NG)

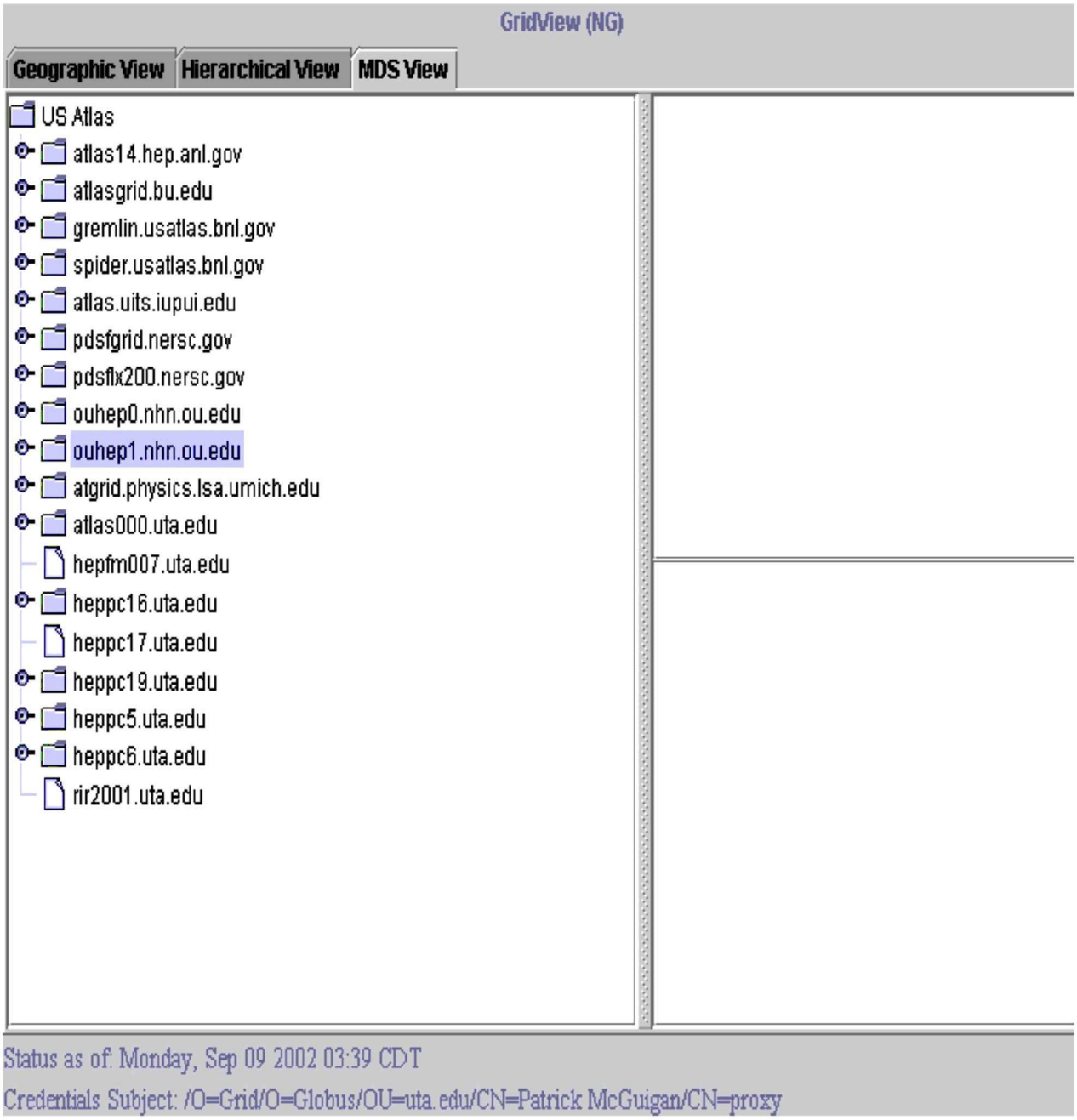

GridView (NG)
Geographic View
Hierarchical View
MDS View
PASS
PASS
PASS
MIXED
MIXED
MIXED
MIXED
MIXED
Status as of: Monday, Sep 09 2002 03:39 CDT
Credentials Subject: /O=Grid/O=Globus/OU=uta.edu/CN=Patrick McGuigan/CN=proxy

# GridView - Tree View

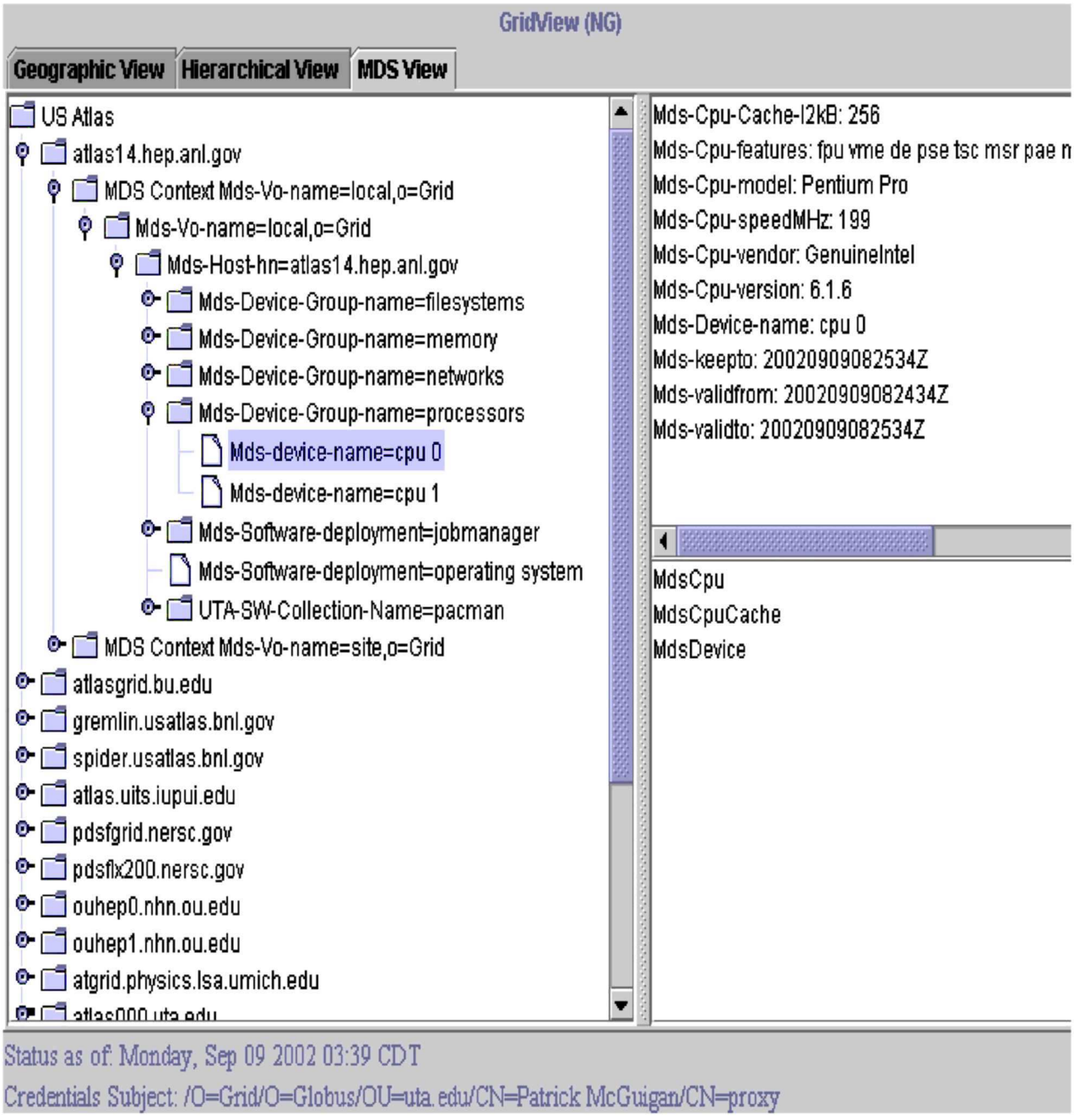
# Test Results



# MDS View

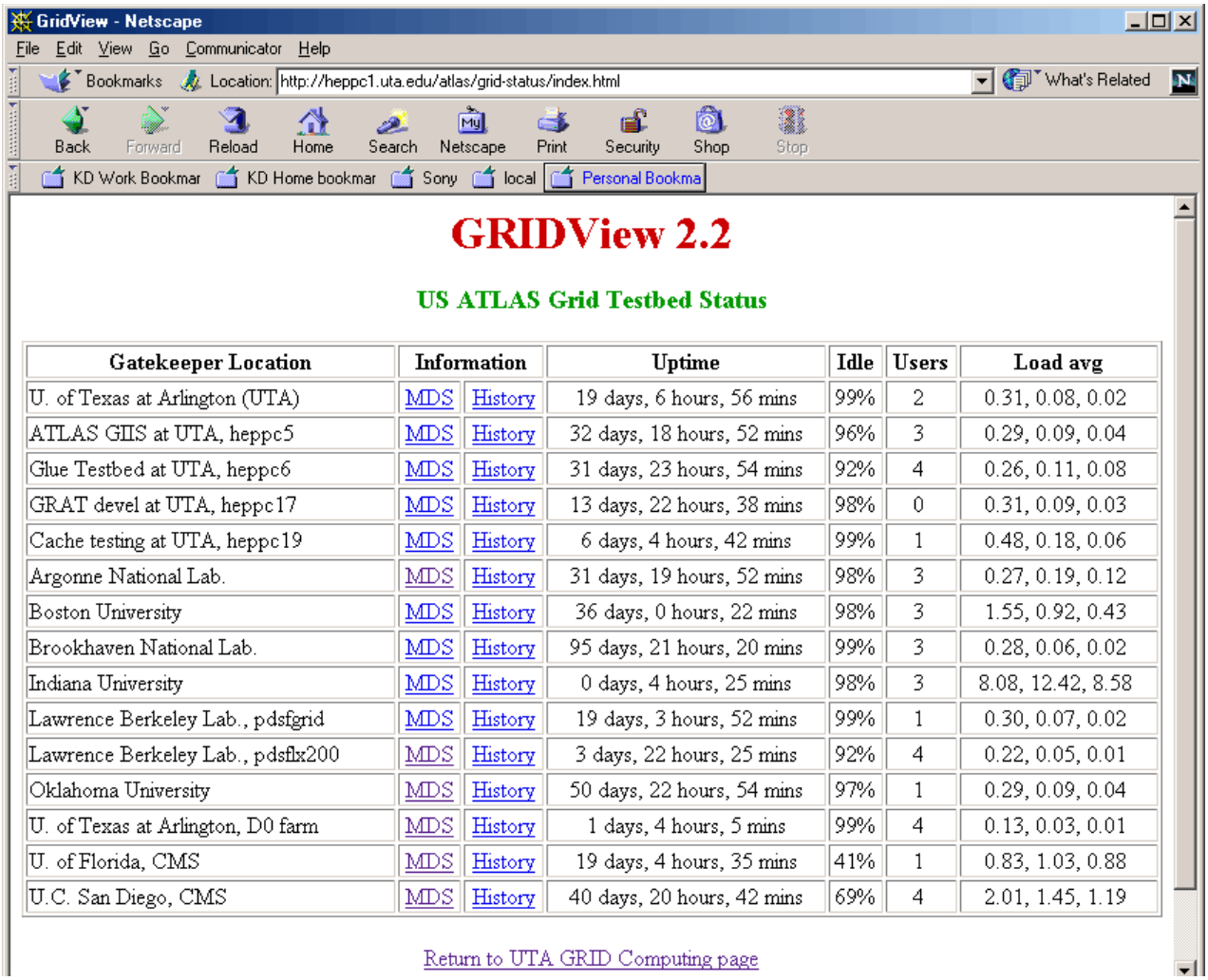

GridView (NG)
Geographic View
Hierarchical View
MDS View
US Atlas
atlas14.hep.anl.gov
atlasgrid.bu.edu
gremlin.usatlas.bnl.gov
spider.usatlas.bnl.gov
atlas.uits.iupui.edu
pdsfgrid.nersc.gov
pdsflx200.nersc.gov
ouhep0.nhn.ou.edu
ouhep1.nhn.ou.edu
atgrid.physics.lsa.umich.edu
atlas000.uta.edu
hepfm007.uta.edu
heppc16.uta.edu
heppc17.uta.edu
heppc19.uta.edu
heppc5.uta.edu
heppc6.uta.edu
rir2001.uta.edu
Status as of: Monday, Sep 09 2002 03:39 CDT
Credentials Subject: /O=Grid/O=Globus/OU=uta.edu/CN=Patrick McGuigan/CN=proxy

# MDS Browser


GridView (NG)
Geographic View
Hierarchical View
MDS View
US Atlas
atlas14.hep.anl.gov
MDS Context Mds-Vo-name=local,o=Grid
Mds-Vo-name=local,o=Grid
Mds-Host-hn=atlas14.hep.anl.gov
Mds-Device-Group-name=filesystems
Mds-Device-Group-name=memory
Mds-Device-Group-name=networks
Mds-Device-Group-name=processors
Mds-device-name=cpu 0
Mds-device-name=cpu 1
Mds-Software-deployment=jobmanager
Mds-Software-deployment=operating system
UTA-SW-Collection-Name=pacman
MDS Context Mds-Vo-name=site,o=Grid
atlasgrid.bu.edu
gremlin.usatlas.bnl.gov
spider.usatlas.bnl.gov
atlas.uits.iupui.edu
pdsfgrid.nersc.gov
pdsflx200.nersc.gov
ouhep0.nhn.ou.edu
ouhep1.nhn.ou.edu
atgrid.physics.lsa.umich.edu
Mds-Cpu-Cache-l2kB: 256
Mds-Cpu-features: fpu vme de pse tsc msr pae n
Mds-Cpu-model: Pentium Pro
Mds-Cpu-speedMHz: 199
Mds-Cpu-vendor: GenuineIntel
Mds-Cpu-version: 6.1.6
Mds-Device-name: cpu 0
Mds-keepto: 20020909082534Z
Mds-validfrom: 20020909082434Z
Mds-validto: 20020909082534Z
MdsCpu
MdsCpuCache
MdsDevice
Status as of: Monday, Sep 09 2002 03:39 CDT
Credentials Subject: /O=Grid/O=Globus/OU=uta.edu/CN=Patrick McGuigan/CN=proxy

# GridView - Old Text Version

- **Simple visualization tool using Globus Toolkit**
  - First native Globus application for ATLAS grid (March 2001)
  - Collects information using Globus tools. Archival information is stored in MySQL server on a different machine. Data published through web server on a third machine.
- http://heppc1.uta.edu/atlas/grid-status/index.html




## GRIDView 2.2

### US ATLAS Grid Testbed Status

| Gatekeeper Location | Information | | Uptime | Idle | Users | Load avg |
|---|---|---|---|---|---|---|
| U. of Texas at Arlington (UTA) | MDS | History | 19 days, 6 hours, 56 mins | 99% | 2 | 0.31, 0.08, 0.02 |
| ATLAS GIIS at UTA, heppc5 | MDS | History | 32 days, 18 hours, 52 mins | 96% | 3 | 0.29, 0.09, 0.04 |
| Glue Testbed at UTA, heppc6 | MDS | History | 31 days, 23 hours, 54 mins | 92% | 4 | 0.26, 0.11, 0.08 |
| GRAT devel at UTA, heppc17 | MDS | History | 13 days, 22 hours, 38 mins | 98% | 0 | 0.31, 0.09, 0.03 |
| Cache testing at UTA, heppc19 | MDS | History | 6 days, 4 hours, 42 mins | 99% | 1 | 0.48, 0.18, 0.06 |
| Argonne National Lab. | MDS | History | 31 days, 19 hours, 52 mins | 98% | 3 | 0.27, 0.19, 0.12 |
| Boston University | MDS | History | 36 days, 0 hours, 22 mins | 98% | 3 | 1.55, 0.92, 0.43 |
| Brookhaven National Lab. | MDS | History | 95 days, 21 hours, 20 mins | 99% | 3 | 0.28, 0.06, 0.02 |
| Indiana University | MDS | History | 0 days, 4 hours, 25 mins | 98% | 3 | 8.08, 12.42, 8.58 |
| Lawrence Berkeley Lab., pdsfgrid | MDS | History | 19 days, 3 hours, 52 mins | 99% | 1 | 0.30, 0.07, 0.02 |
| Lawrence Berkeley Lab., pdsflx200 | MDS | History | 3 days, 22 hours, 25 mins | 92% | 4 | 0.22, 0.05, 0.01 |
| Oklahoma University | MDS | History | 50 days, 22 hours, 54 mins | 97% | 1 | 0.29, 0.09, 0.04 |
| U. of Texas at Arlington, D0 farm | MDS | History | 1 days, 4 hours, 5 mins | 99% | 4 | 0.13, 0.03, 0.01 |
| U. of Florida, CMS | MDS | History | 19 days, 4 hours, 35 mins | 41% | 1 | 0.83, 1.03, 0.88 |
| U.C. San Diego, CMS | MDS | History | 40 days, 20 hours, 42 mins | 69% | 4 | 2.01, 1.45, 1.19 |

Return to UTA GRID Computing page

# Testbed Tools

- Many tools developed by the U.S. ATLAS testbed group during past year
- GridView - simple tool to monitor status of testbed Kaushik De, Patrick McGuigan
- Gripe - unified user accounts Rob Gardner
- Magda - MAnager for Grid DAta Torre Wenaus, Wensheng Deng (see Gardner & Wenaus talks)
- Pacman - package management and distribution tool Saul Youssef
  - Being widely used or adopted by iVDGL VDT, Ganga, and others (see Gardner talk)
- Grappa - web portal using active notebook technology Shava Smallen (see Gardner talk)
- GRAT - GRid Application Toolkit
- Gridsearcher - MDS browser Jennifer Schopf
- GridExpert - Knowledge Database Mark Sosebee
- VO Toolkit - Site AA Rich Baker (see Baker talk)
- ...

# Integration!!

- Coordination with other grid efforts and software developers - very difficult task!
- Project centric:
  - GriPhyN/iVDGL - Rob Gardner
  - PPDG - Torre Wenaus
  - EDG - Ed May, Jerry Gieraltowski
  - ATLAS/LHCb - Rich Baker
  - ATLAS/CMS - Kaushik De
  - ATLAS/D0 - Jae Yu
- Fabric/Middleware centric:
  - Afs Software installations - Alex Undrus, Shane Canon, Iwona Sakrejda
  - Networking - Shawn McKee, Rob Gardner
  - Virtual and Real Data Management - Wendsheng Deng, Sasha Vaniachin, Pavel Nevski, David Malon, Rob Gardner, Dan Engh, Mike Wilde, Yong Zhao, Shava Smallen
  - Security/Site AA/VO - Rich Baker, Dantong Yu