Bringing Grids to University Campuses

Paul Avery
University of Florida
avery@phys.ufl.edu

**International ICFA Workshop on
HEP, Networking & Digital Divide
Issues for Global e-Science
Daegu, Korea
May 27, 2005**

Examples Discussed Here

- Three campuses, in different states of readiness
  - University of Wisconsin: GLOW
  - University of Michigan: MGRID
  - University of Florida: UF Research Grid
- Not complete, by any means
  - Goal is to illustrate factors that go into creating campus Grid facilities

Grid Laboratory of Wisconsin

- 2003 Initiative funded by NSF/UW: Six GLOW Sites
  - Computational Genomics, Chemistry
  - Amanda, Ice-cube, Physics/Space Science
  - High Energy Physics/CMS, Physics
  - Materials by Design, Chemical Engineering
  - Radiation Therapy, Medical Physics
  - Computer Science
- Deployed in two Phases

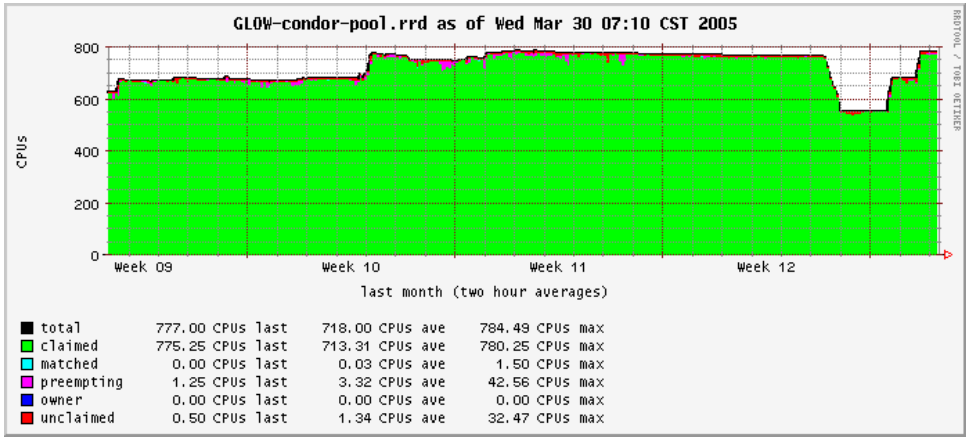
http://www.cs.wisc.edu/condor/glow/

Condor/GLOW Ideas

- Exploit commodity hardware for high throughput computing
  - The base hardware is the same at all sites
  - Local configuration optimization as needed (e.g., CPU vs storage)
  - Must meet global requirements (very similar configurations now)
- Managed locally at 6 sites
  - Shared globally across all sites
  - Higher priority for local jobs

GLOW Deployment

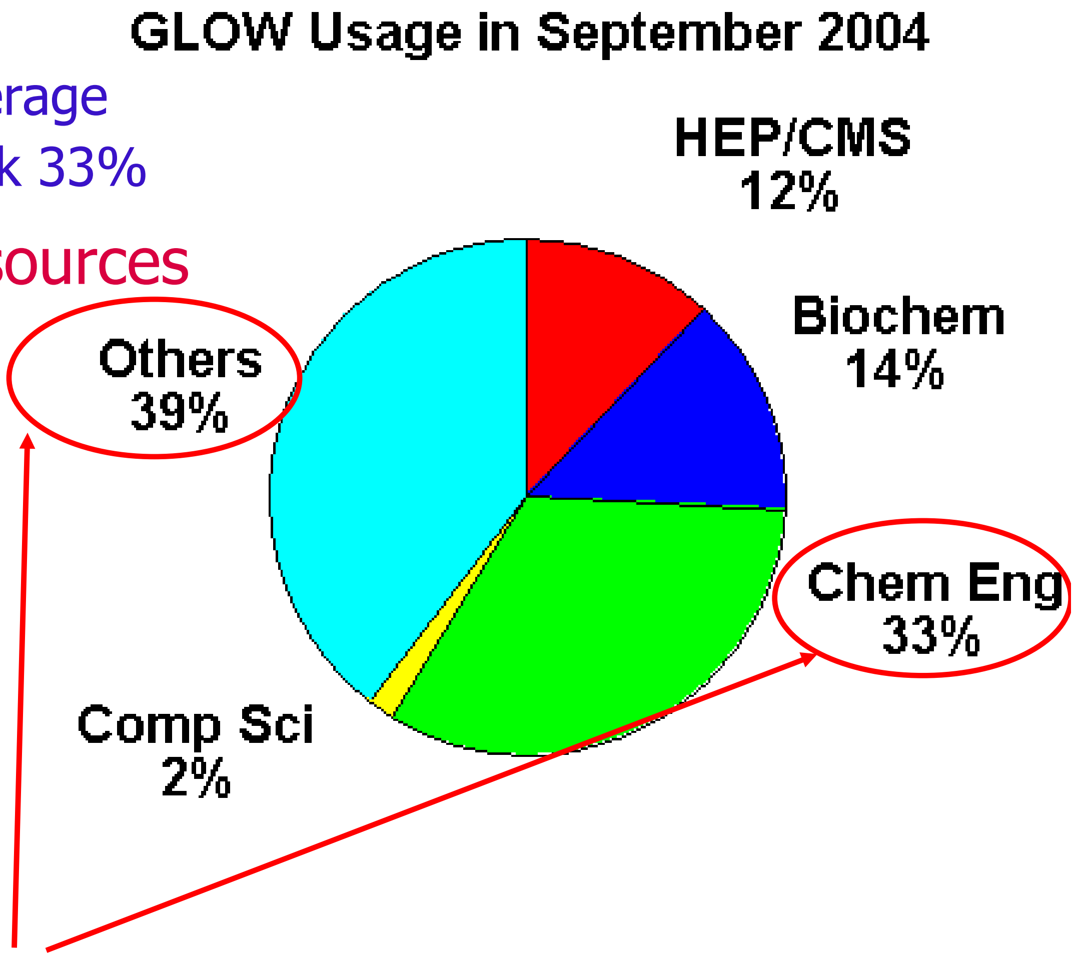
- GLOW Phase-I and II are commissioned
- CPU
  - 66 nodes each @ ChemE, CS, LMCG, MedPhys
  - 60 nodes @ Physics
  - 30 nodes @ IceCube
  - 50 extra nodes @ CS (ATLAS)
  - Total CPU: ~800
- Storage
  - Head nodes @ at all sites
  - 45 TB each @ CS and Physics
  - Total storage: ~ 100 TB
- GLOW resources used at ~100% level
  - Key is having multiple user groups



■ total	777.00 CPUs last	718.00 CPUs ave	784.49 CPUs max
🟩 claimed	775.25 CPUs last	713.31 CPUs ave	780.25 CPUs max
■ matched	0.00 CPUs last	0.03 CPUs ave	1.50 CPUs max
■ preempting	1.25 CPUs last	3.32 CPUs ave	42.56 CPUs max
🟦 owner	0.00 CPUs last	0.00 CPUs ave	0.00 CPUs max
🟥 unclaimed	0.50 CPUs last	1.34 CPUs ave	32.47 CPUs max

Resource Sharing in GLOW

- Six GLOW sites
  - Equal priority ⇒ 17% average
  - Chemical Engineering took 33%
- Others scavenge idle resources
  - Yet, they got 39%

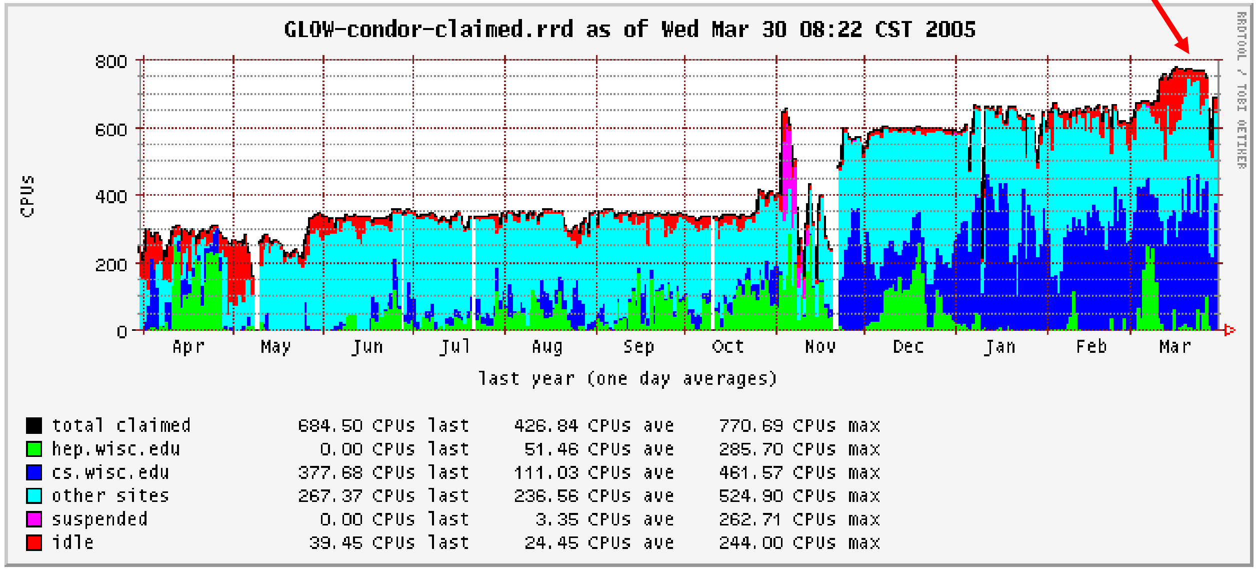


Efficient users can realize much more than they put in

GLOW Usage: Highly Efficient

- CS + Guests
  - Largest user, many cycles delivered to guests
- ChemE
  - Largest community
- HEP/CMS
  - Production for collaboration, analysis for local physicists
- LMCG
  - Standard Universe
- Medical Physics
  - MPI jobs
- IceCube
  - Simulations

~800 CPUs



■ total claimed	684.50 CPUs last	426.84 CPUs ave	770.69 CPUs max
🟩 hep.wisc.edu	0.00 CPUs last	51.46 CPUs ave	285.70 CPUs max
🟦 cs.wisc.edu	377.68 CPUs last	111.03 CPUs ave	461.57 CPUs max
□ other sites	267.37 CPUs last	236.56 CPUs ave	524.90 CPUs max
🟪 suspended	0.00 CPUs last	3.35 CPUs ave	262.71 CPUs max
🟥 idle	39.45 CPUs last	24.45 CPUs ave	244.00 CPUs max

Adding New GLOW Members

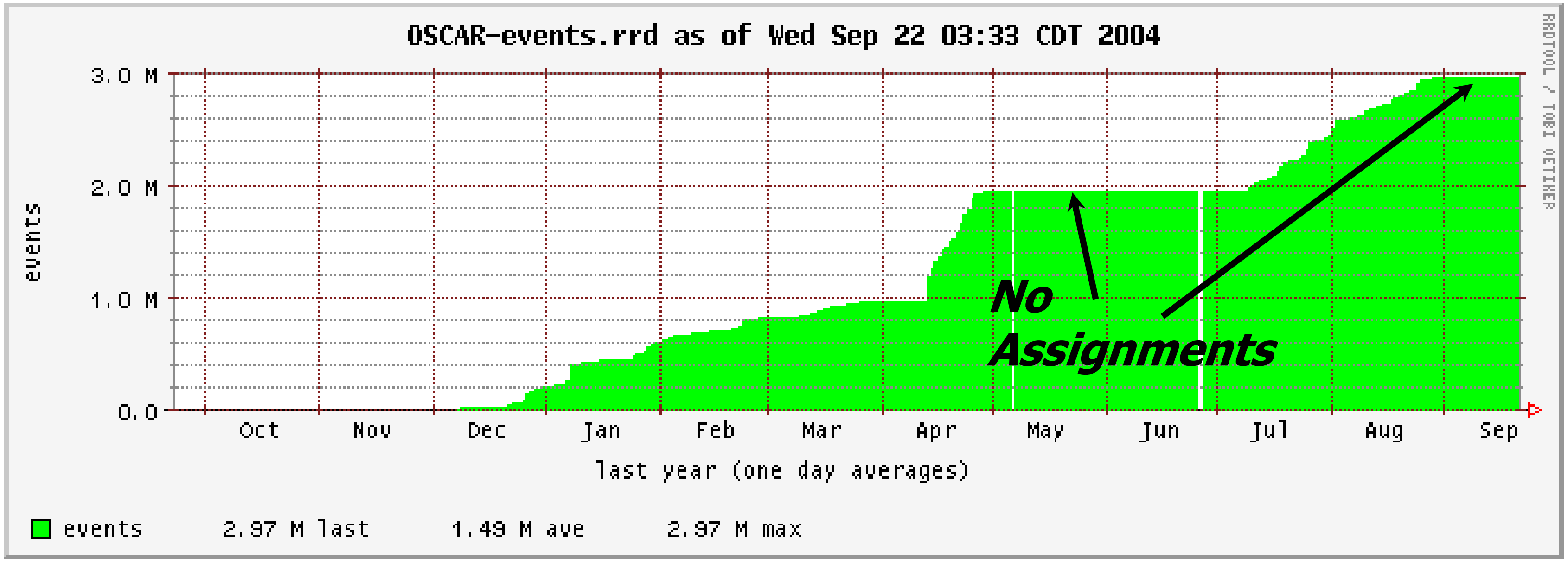
- Proposed minimum involvement
  - One rack with about 50 CPUs
- Identified system support person who joins GLOW-tech
- PI joins the GLOW-exec
- Adhere to current GLOW policies
- Sponsored by existing GLOW members
  - ATLAS group and Condensed matter group were proposed by CMS and CS, and were accepted as new members
    - ATLAS using 50% of GLOW cycles (housed @ CS)
    - New machines of CM Physics group being commissioned
  - Expressions of interest from other groups

GLOW & Condor Development

- GLOW presents CS researchers with an ideal laboratory
  - Real users with diverse requirements
  - Early commissioning and stress testing of new Condor releases in an environment controlled by Condor team
  - Results in robust releases for world-wide Condor deployment
- New features in Condor Middleware (examples)
  - Group wise or hierarchical priority setting
  - Rapid-response with large resources for short periods of time for high priority interrupts
  - Hibernating shadow jobs instead of total preemption
  - MPI use (Medical Physics)
  - Condor-G (High Energy Physics)

OSCAR Simulation on Condor/GLOW

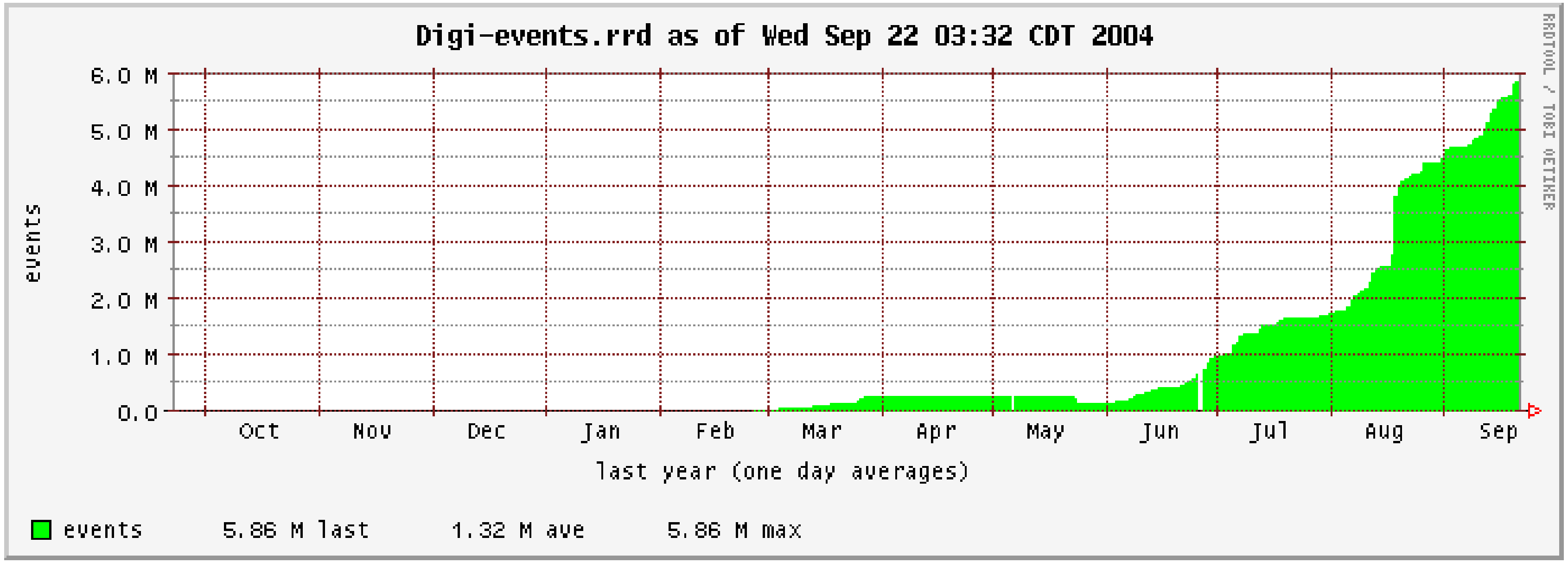
- OSCAR - Simulation using Geant4
  - Runs in Vanilla Universe only (no checkpointing possible)
  - Poor efficiency because of lack of checkpointing
  - Application level checkpointing not in production (yet)



CMS Reconstruction on Condor/GLOW

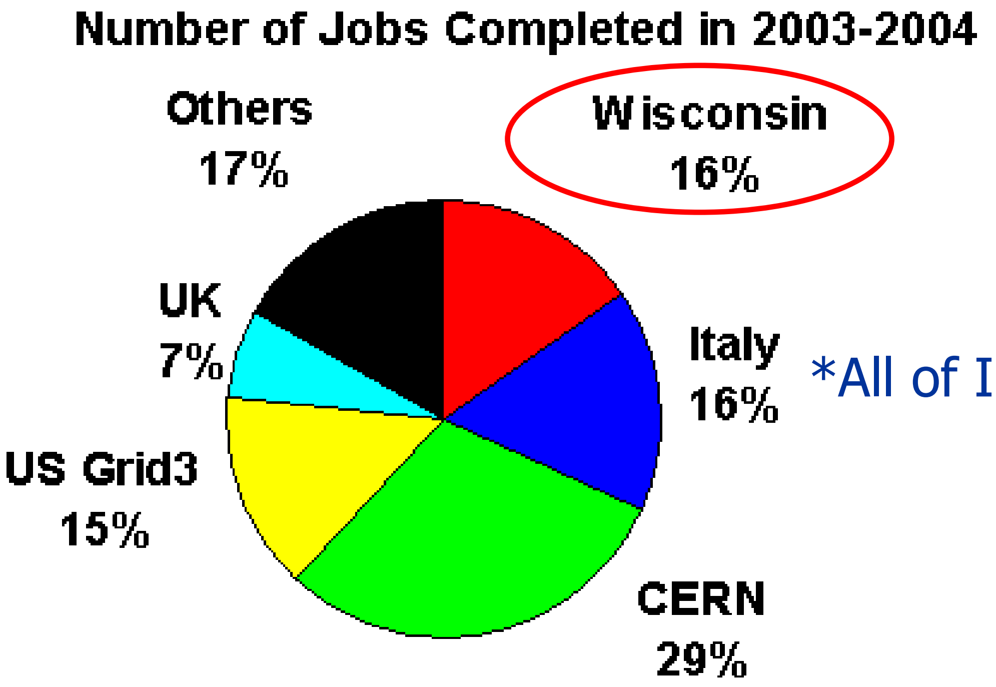
- ORCA - Digitization
  - Vanilla Universe only (no checkpointing)
- IO Intensive
  - Used Fermilab/DESY dCache system
  - Automatic replication of frequently accessed “pileup” events

2004 production



CMS Work Done on Condor/GLOW

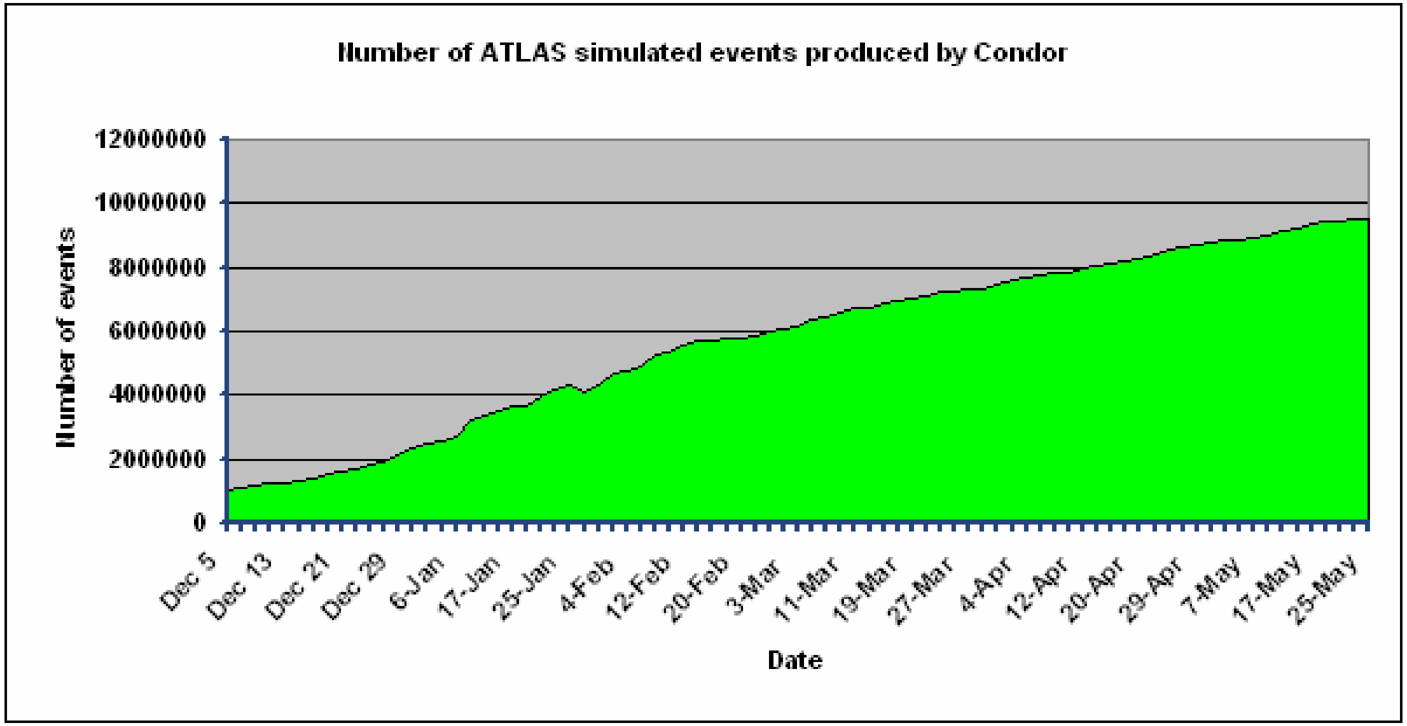
- UW Condor/GLOW was top source for CMS production
  - Largest single institution excluding DC04 DST production at CERN



*All of INFN

ATLAS Simulations at GLOW

~9.5M events generated in 2004



MGRID at Michigan

- MGRID
  - Michigan Grid Research and Infrastructure Development
  - Develop, deploy, and sustain an institutional grid at Michigan
  - Group started in 2002 with initial U Michigan funding
- Many groups across the University participate
  - Compute/data/network-intensive research grants
  - ATLAS, NPACI, NEESGrid, Visible Human, NFSv4, NMI

http://www.mgrid.umich.edu

MGRID Center

- Central core of technical staff (3FTEs, new hires)
- Faculty and staff from participating units
- Exec. committee from participating units & provost office
- Collaborative grid research and development with technical staff from participating units

MGrid Research Project Partners

- College of LS&A (Physics) (www.lsa.umich.edu)
- Center for Information Technology Intergration (www.citi.umich.edu)
- Michigan Center for BioInformatics(www.ctaalliance.org)
- Visible Human Project (vhp.med.umich.edu)
- Center for Advanced Computing (cac.engin.umich.edu)
- Mental Health Research Institute (www.med.umich.edu/mhri)
- ITCom (www.itcom.itd.umich.edu)
- School of Information (si.umich.edu)

MGRID: Goals

- For participating units
  - Knowledge, support and framework for deploying Grid technologies
  - Exploitation of Grid resources both on campus and beyond
  - A context for the University to invest in computing resources
- Provide test bench for existing, emerging Grid technologies
- Coordinate activities within the national Grid community
  - GGF, GlobusWorld, etc
- Make significant contributions to general grid problems
  - Sharing resources among multiple VOs
  - Network monitoring and QoS issues for grids
  - Integration of middleware with domain specific applications
  - Grid filesystems

MGRID Authentication

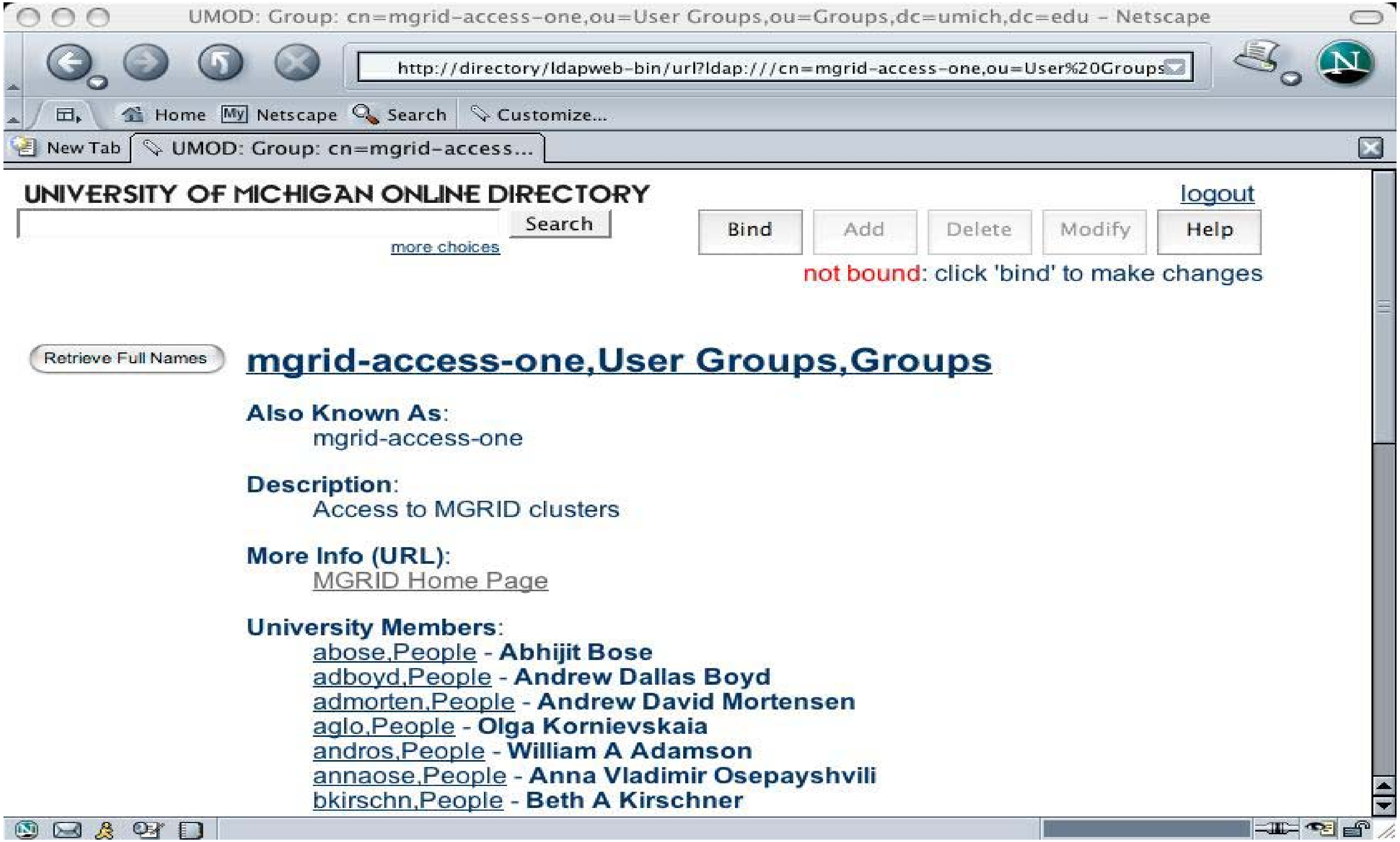
- Developed a KX509 module that bridges two technologies
  - Globus public key cryptography (X509 certificates)
  - UM Kerberos user authentication
- MGRID provides step-by-step instructions on web site
  - "How to Grid-Enable Your Browser"

MGRID Authorization

- MGRID uses Walden: fine-grained authorization engine
  - Leveraging open-source XACML implementation from Sun
- Walden allows interesting granularity of authorization
  - Definition of authorization user groups
  - Each group has a different level of authority to run a job
  - Authority level depends on conditions (job queue, time of day, CPU load, …)
- Resource owners still have complete control over user membership within these groups

MFRID Authorization Groups

- Authorization groups defined through UM Online Directory, or viaMGRID Directory for external users



UMOD: Group: cn=mgrid-access-one,ou=User Groups,ou=Groups,dc=umich,dc=edu - Netscape
http://directory/ldapweb-bin/url?ldap:///cn=mgrid-access-one,ou=User%20Groups
Home
Netscape
Search
Customize...
New Tab
UMOD: Group: cn=mgrid-access...
Search
more choices
logout
Bind
Add
Delete
Modify
Help
Retrieve Full Names

UNIVERSITY OF MICHIGAN ONLINE DIRECTORY

not bound: click 'bind' to make changes

mgrid-access-one,User Groups,Groups

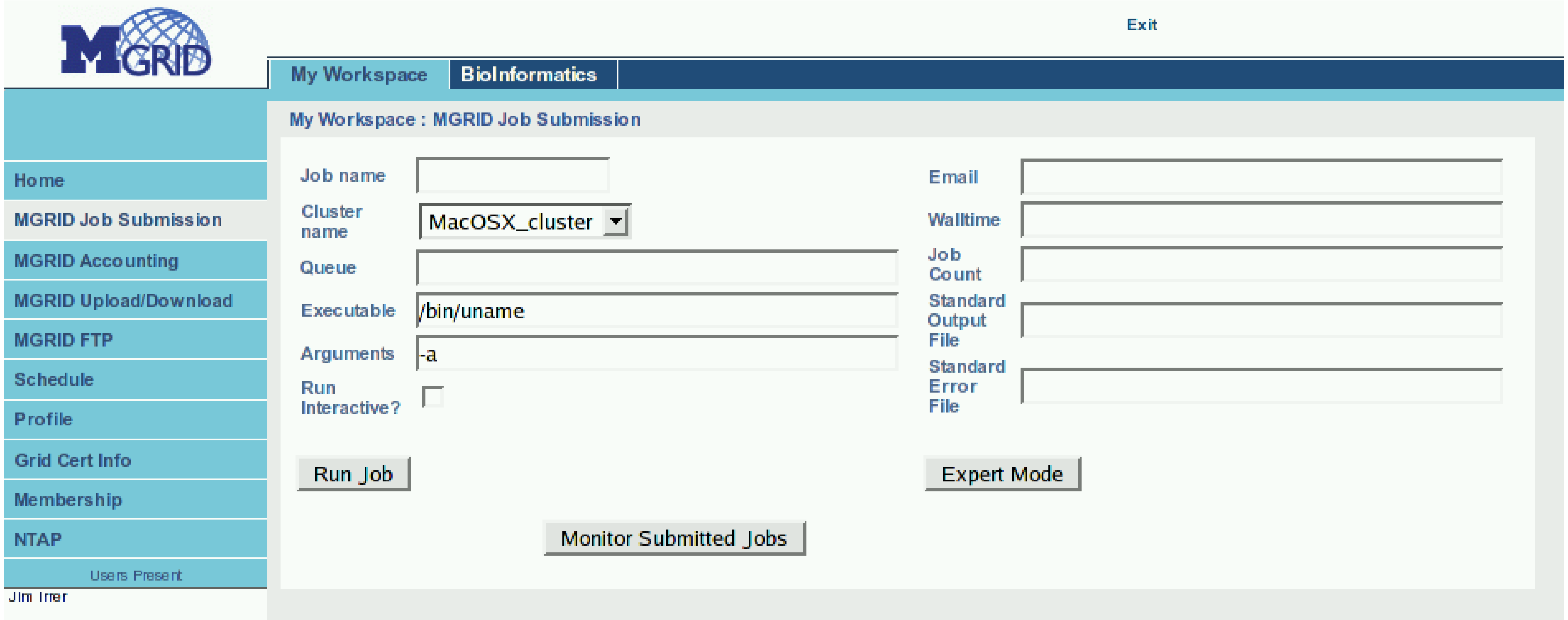
Also Known As:
mgrid-access-one

Description:
Access to MGRID clusters

More Info (URL):
MGRID Home Page

University Members:
abose,People - **Abhijit Bose**
adboyd,People - **Andrew Dallas Boyd**
admorten,People - **Andrew David Mortensen**
aglo,People - **Olga Kornievskaia**
andros,People - **William A Adamson**
annaose,People - **Anna Vladimir Osepayshvili**
bkirschn,People - **Beth A Kirschner**

MGRID Job Portal



My Workspace : MGRID Job Submission

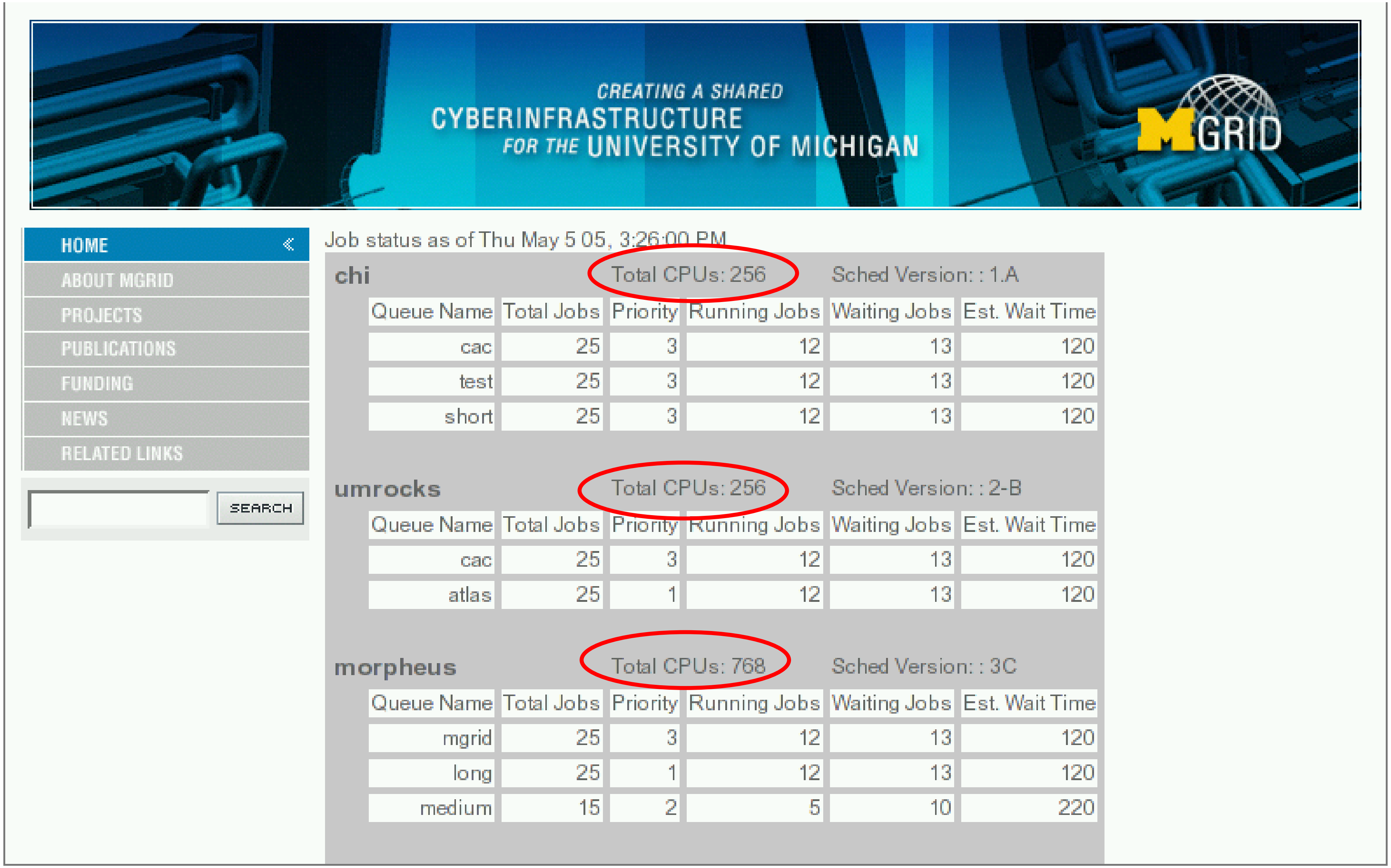
Cluster name: MacOSX_cluster

Executable: /bin/uname

Arguments: -a

Run Interactive? □

MGRID Job Status



Job status as of Thu May 5 05, 3:26:00 PM

chi Total CPUs: 256 Sched Version: : 1.A

Queue Name	Total Jobs	Priority	Running Jobs	Waiting Jobs	Est. Wait Time
cac	25	3	12	13	120
test	25	3	12	13	120
short	25	3	12	13	120

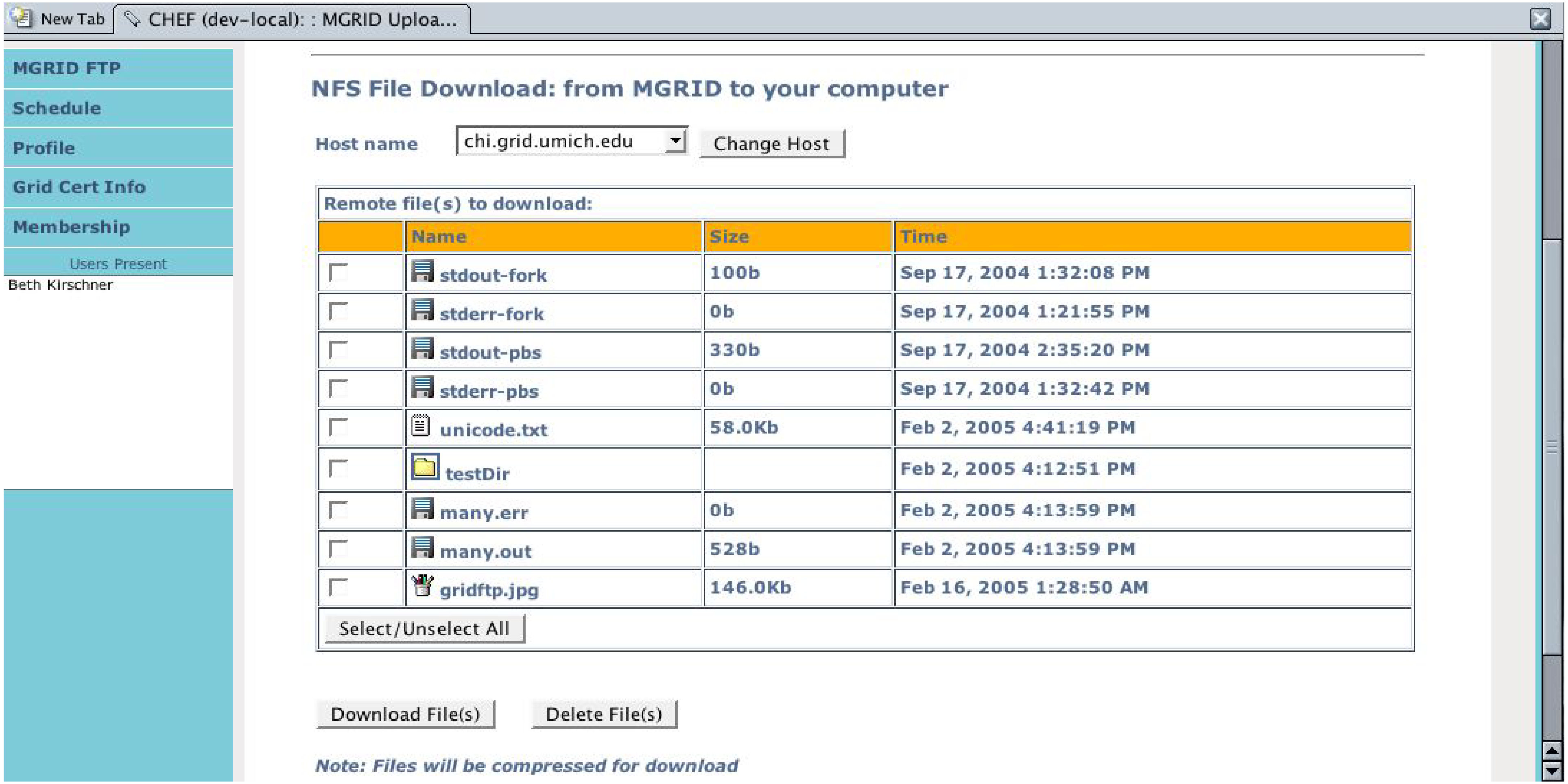
umrocks Total CPUs: 256 Sched Version: : 2-B

Queue Name	Total Jobs	Priority	Running Jobs	Waiting Jobs	Est. Wait Time
cac	25	3	12	13	120
atlas	25	1	12	13	120

morpheus Total CPUs: 768 Sched Version: : 3C

Queue Name	Total Jobs	Priority	Running Jobs	Waiting Jobs	Est. Wait Time
mgrid	25	3	12	13	120
long	25	1	12	13	120
medium	15	2	5	10	220

MGRID File Upload/Download



Beth Kirschner

NFS File Download: from MGRID to your computer

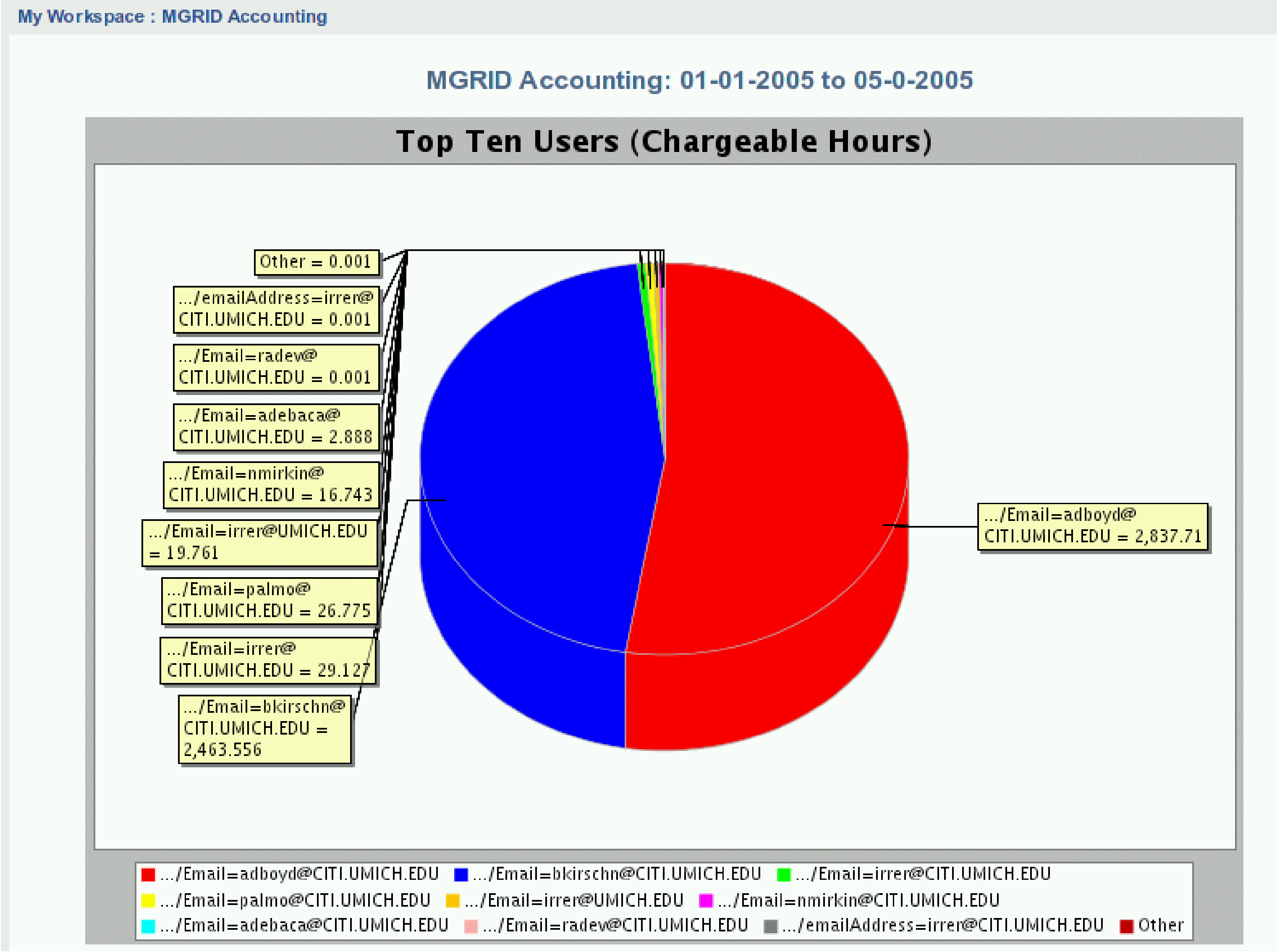
Host name chi.grid.umich.edu

Remote file(s) to download:

	Name	Size	Time
□	stdout-fork	100b	Sep 17, 2004 1:32:08 PM
□	stderr-fork	0b	Sep 17, 2004 1:21:55 PM
□	stdout-pbs	330b	Sep 17, 2004 2:35:20 PM
□	stderr-pbs	0b	Sep 17, 2004 1:32:42 PM
□	unicode.txt	58.0Kb	Feb 2, 2005 4:41:19 PM
□	testDir		Feb 2, 2005 4:12:51 PM
□	many.err	0b	Feb 2, 2005 4:13:59 PM
□	many.out	528b	Feb 2, 2005 4:13:59 PM
□	gridftp.jpg	146.0Kb	Feb 16, 2005 1:28:50 AM

Note: Files will be compressed for download

Major MGRID Users (Example)



My Workspace : MGRID Accounting

MGRID Accounting: 01-01-2005 to 05-0-2005

Top Ten Users (Chargeable Hours)

Other = 0.001
.../emailAddress=irrer@CITI.UMICH.EDU = 0.001
.../Email=radev@CITI.UMICH.EDU = 0.001
.../Email=adebaca@CITI.UMICH.EDU = 2.888
.../Email=nmirkin@CITI.UMICH.EDU = 16.743
.../Email=irrer@UMICH.EDU = 19.761
.../Email=palmo@CITI.UMICH.EDU = 26.775
.../Email=irrer@CITI.UMICH.EDU = 29.127
.../Email=bkirschn@CITI.UMICH.EDU = 2,463.556
.../Email=adboyd@CITI.UMICH.EDU = 2,837.71
🟥 .../Email=adboyd@CITI.UMICH.EDU
🟦 .../Email=bkirschn@CITI.UMICH.EDU
🟩 .../Email=irrer@CITI.UMICH.EDU
🟨 .../Email=palmo@CITI.UMICH.EDU
🟧 .../Email=irrer@UMICH.EDU
🟪 .../Email=nmirkin@CITI.UMICH.EDU
■ .../Email=adebaca@CITI.UMICH.EDU
■ .../Email=radev@CITI.UMICH.EDU
■ .../emailAddress=irrer@CITI.UMICH.EDU
■ Other

University of Florida Research Grid

- High Performance Computing Committee: April 2001
  - Created by Provost & VP for Research
  - Currently has 16 members from around campus
- Study in 2001-2002
  - UF Strength: Faculty expertise and reputation in HPC
  - UF Weakness: Infrastructure lags well behind AAU public peers
- Major focus
  - Create campus Research Grid with HPC Center as kernel
  - Expand research in HPC-enabled applications areas
  - Expand research in HPC infrastructure research
  - Enable new collaborations, visibility, external funding, etc.

http://www.hpc.ufl.edu/CampusGrid/

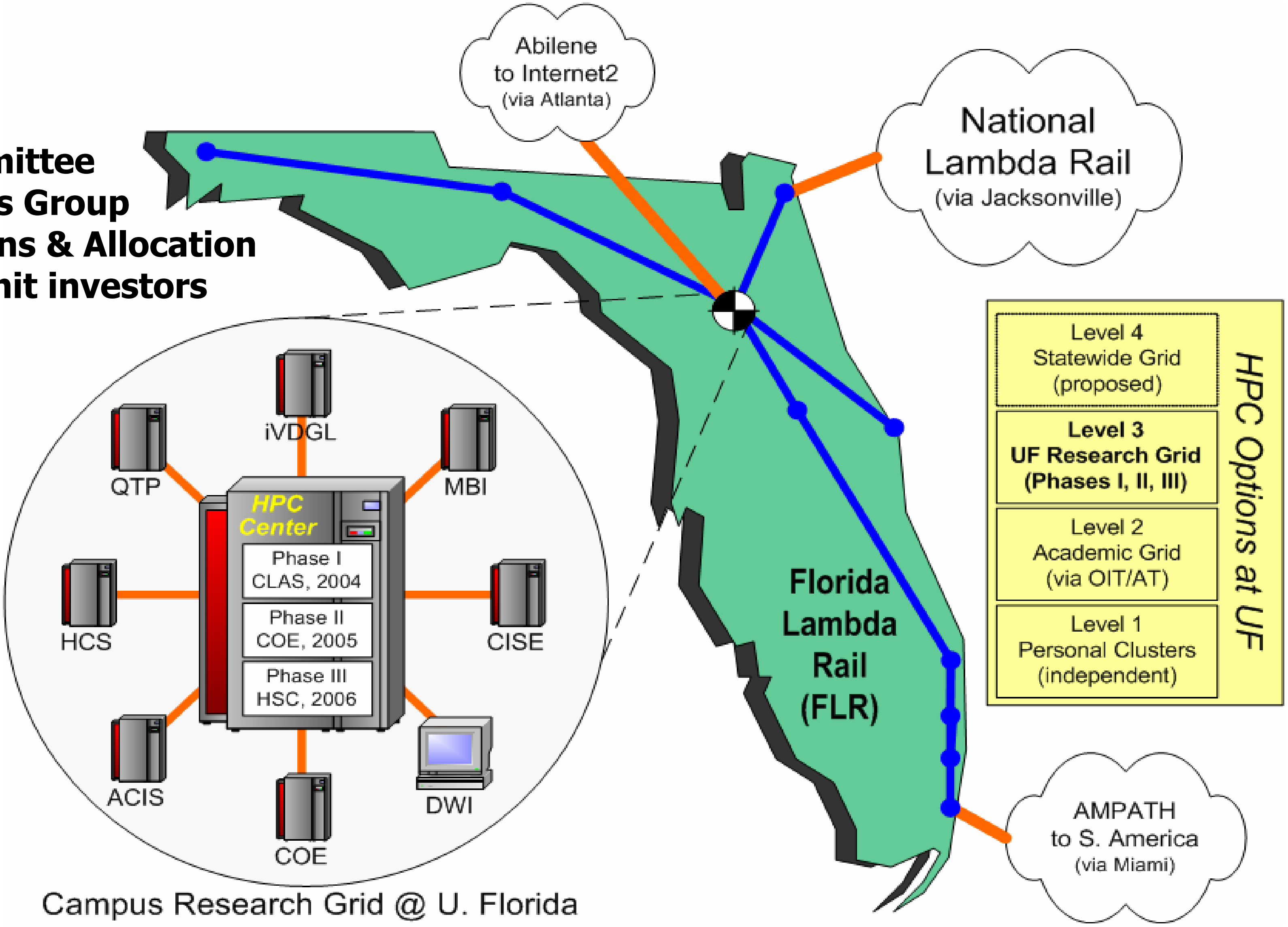
UF Grid Strategy

- A campus-wide, distributed HPC facility
  - Multiple facilities, organization, resource sharing
  - Staff, seminars, training
- Faculty-led, research-driven, investor-oriented approach
  - With administrative cost-matching & buy-in by key vendors
- Build basis for new multidisciplinary collaborations in HPC
  - HPC as a key common denominator for multidisciplinary research
- Expand research opportunities for broad range of faculty
  - Including those already HPC-savvy and those new to HPC
- Build HPC Grid facility in 3 phases
  - Phase I: Investment by College of Arts & Sciences (in operation)
  - Phase II: Investment by College of Engineering (in develpment)
  - Phase III: Investment by Health Science Center (in 2006)

UF HPC Center and Research Grid

Oversight

- **HPC Committee**
- **Operations Group**
- **Applications & Allocation**
- **Faculty/unit investors**



Phase I (Coll. of Arts & Sciences Focus)

- Physics
  - $200K for equipment investment
- College of Arts and Sciences
  - $100K for equipment investment, $70K/yr systems engineer
- Provost's office
  - $300K matching for equipment investment
  - ~$80K/yr Sr. HPC systems engineer
  - ~$75K for physics computer room renovation
  - ~$10K for an open account for various HPC Center supplies
- Now deployed (see next slides)

Phase I Facility (Fall 2004)

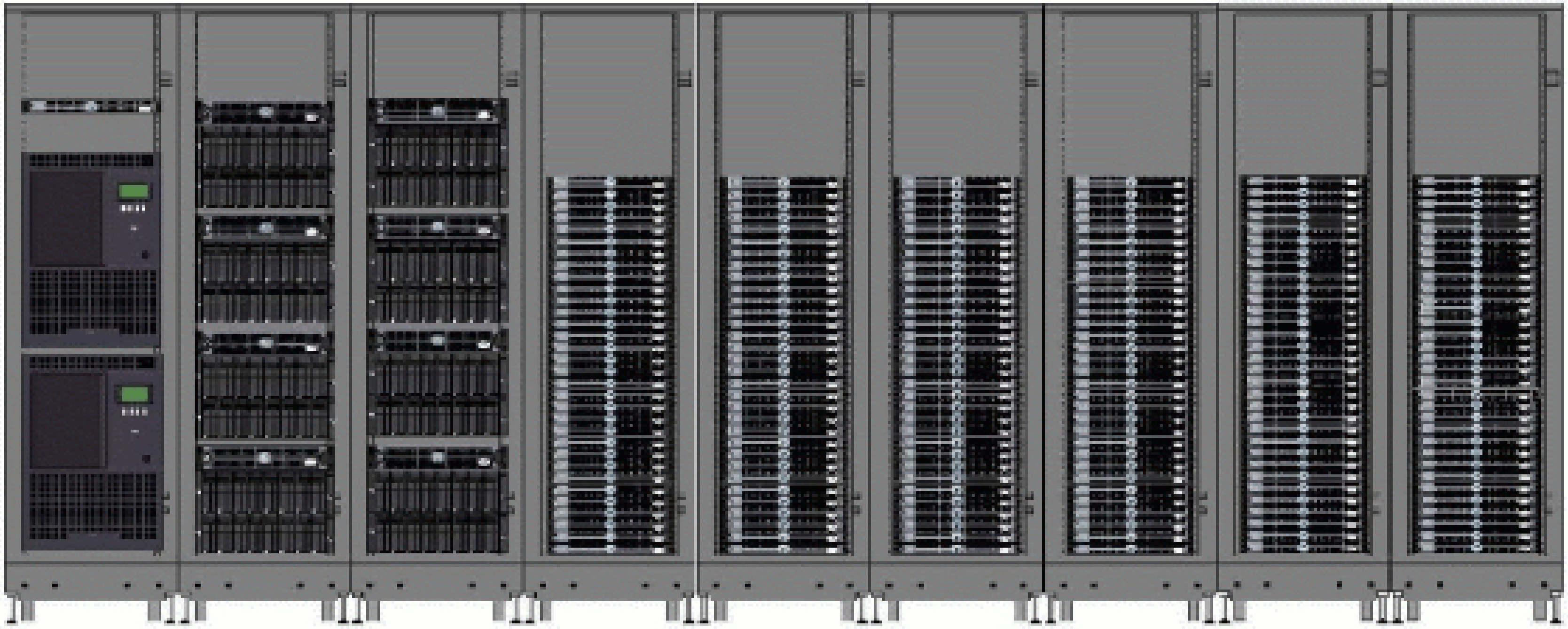
- 200-node cluster of dual-Xeon machines
  - 192 compute nodes (dual 2.8 GHz, 2GB memory, 74 GB disk)
  - 8 I/O nodes (32 of storage in SCSI RAID)
  - Tape unit for some backup
  - 3 years of hardware maintenance
- 1.325 TFLOPS (#221 on Top500)

TOP500

This Site is ranked

NO. 221 in the TOP500 List

published Nov/2004



–

Phase I HPC Use

- Early period (2-3 months) of severe underuse
  - Not “discovered”
  - Lack of documentation
  - Need for early adopters
- Currently enjoying high level of use (> 90%)
  - CMS production simulations
  - Other Physics
  - Quantum Chemistry
  - Other chemistry
  - Health sciences
  - Several engineering apps

Phase I HPC Use (cont)

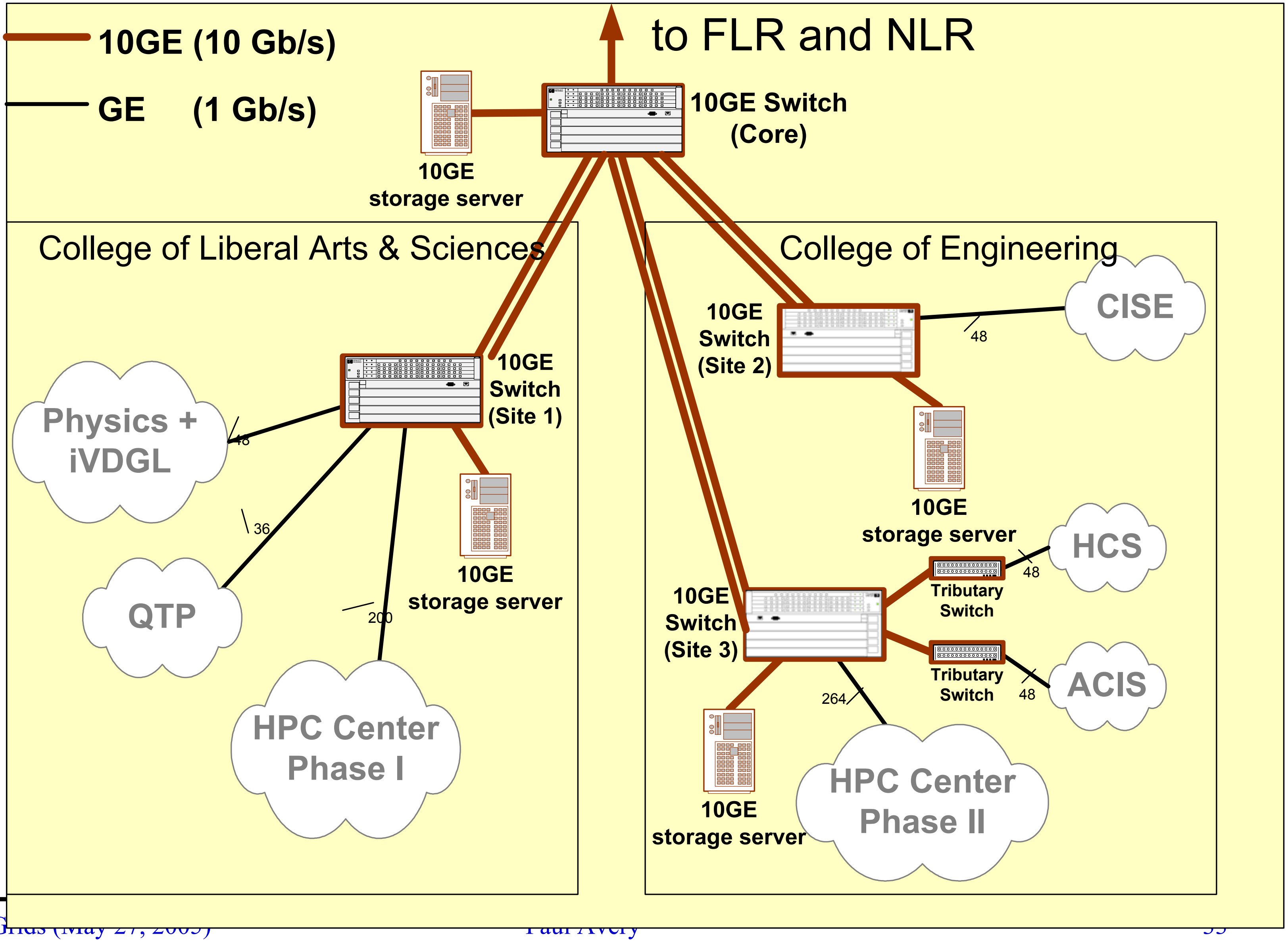
- Still primitive, in many respects
  - Insufficient monitoring & display
  - No accounting yet
  - Few services (compared to Condor, MGRID)
- Job portals
  - PBS is currently main job portal
  - New In-VIGO portal being developed (http://invigo.acis.ufl.edu/)
  - Working with TACC (Univ. of Texas) to deploy GridPort
- Plan to leverage tools & services from others
  - Other campuses: GLOW, MGRID, TACC, Buffalo
  - Open Science Grid

New HPC Resources

- Recent NSF/MRI proposal for networking infrastructure
  - $600K: 20 Gb/s network backbone
  - High performance storage (distributed)
- Recent funding of UltraLight and DISUN proposals
  - UltraLight ($700K): Advanced uses for optical networks
  - DISUN ($2.5M): CMS, bring advanced IT to other sciences
- Special vendor relationships
  - Dell, Cisco, Ammasso

UF Research Network (20 Gb/s)

Funded by NSF-MRI grant



Resource Allocation Strategy

- Faculty/unit investors are first preference
  - Top-priority access commensurate with level of investment
  - Shared access to all available resources
- Cost-matching by administration offers many benefits
  - Key resources beyond computation (storage, networks, facilities)
  - Support for broader user base than simply faculty investors
- Economy of scale advantages with broad HPC Initiative
  - HPC vendor competition, strategic relationship, major discounts
  - Facilities savings (computer room space, power, cooling, staff)

Phase II (Engineering Focus)

- Funds being collected now from Engineering faculty
  - Electrical and Computer Engineering
  - Mechanical Engineering
  - Material Sciences
  - Chemical Engineering (possible)
- Matching funds (including machine room & renovations)
  - Engineering departments
  - College of Engineering
  - Provost
- Equipment expected in Phase II facility (Fall 2005)
  - ~400 dual nodes
  - ~100 TB disk
  - High-speed switching fabric
  - (20 Gb/s network backbone)

Phase III (Health Sciences Focus)

- Planning committee formed by HSC in Dec '04
  - Submitting recommendations to HSC administration in May
- Defining HPC needs of Health Science
  - Not only computation; heavy needs in comm. and storage
  - Need support with HPC applications development and use
- Optimistic for major investments in 2006
  - Phase I success & use by Health Sciences are major motivators
  - Process will start in Fall 2005, before Phase II complete