The Virtual Grid Application Development Software (VGrADS) Project

Overview

Ken Kennedy
VGrADS Director
Rice University

http://vgrads.rice.edu/

The VGrADS Team

- VGrADS is an NSF-funded Information Technology Research project

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Dan Reed

RICE

Keith Cooper
Ken Kennedy
Charles Koelbel
Richard Tapia
Linda Torczon

THE UNIVERSITY of TENNESSEE

Jack Dongarra

ISI

Carl Kesselman

UCSD

Fran Berman
Andrew Chien
Henri Casanova

UCSB

Rich Wolski

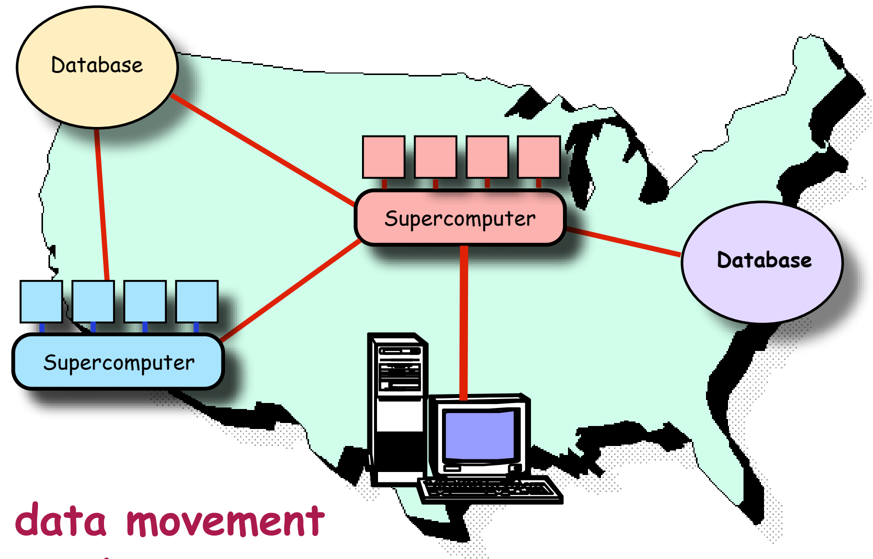
UH

Lennart Johnsson

- Plus many graduate students, postdocs, and technical staff!

Vision: Global Distributed Problem Solving

- Where We Want To Be
  - Transparent Grid computing
    - Submit job
    - Find & schedule resources
    - Execute efficiently
- Where We Are
  - Low-level hand programming
  - Programmer must manage:
    - Heterogeneous resources
    - Scheduling of computation and data movement
    - Fault tolerance and performance adaptation
- What Do We Propose as A Solution?
  - Separate application development from resource management
    - Through an abstraction called the Virtual Grid
  - Provide tools to bridge the gap between conventional and Grid computation
    - Scheduling, resource management, distributed launch, simple programming models, fault tolerance, grid economies

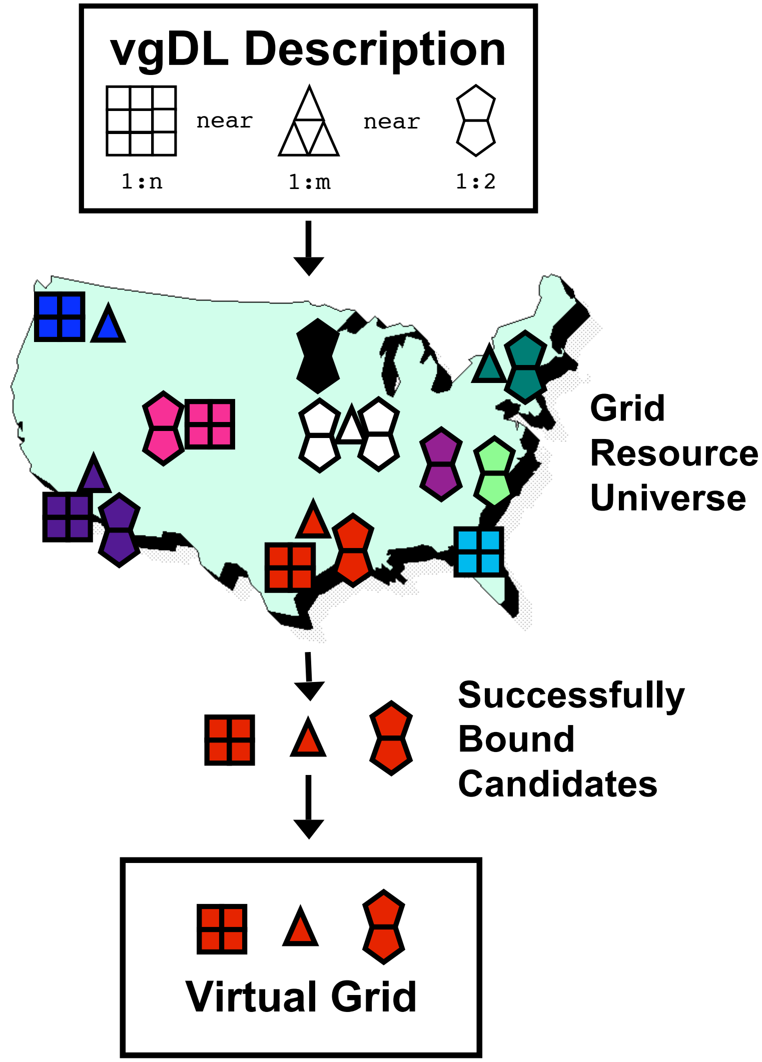


VGrADS Big Ideas

- **Virtualization of Resources**
  - Application specifies required resources in Virtual Grid Definition language (vgDL)
    - Give me a loose bag of 1000 processors, with 1 Gb memory per processor, with the fastest possible processors
    - Give me a tight bag of as many Opterons as possible
  - Virtual Grid Execution System (vgES) produces specific virtual grid matching specification
  - Avoids need for scheduling against the entire space of global resources
- **Generic In-Advance Scheduling of Application Workflows**
  - Application includes performance models for all workflow nodes
    - Performance models automatically constructed
  - Software schedules applications onto virtual Grid, minimizing total makespan
    - Including both computation and data movement times

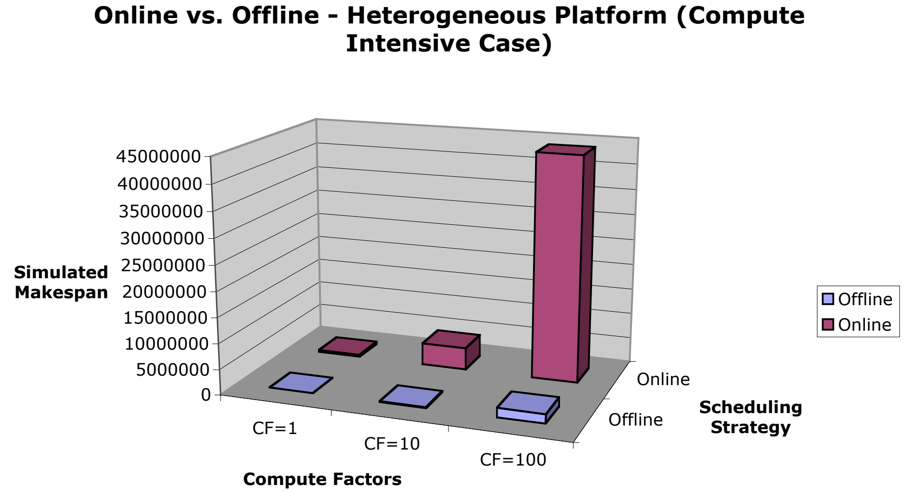
Virtual Grids (VGs)

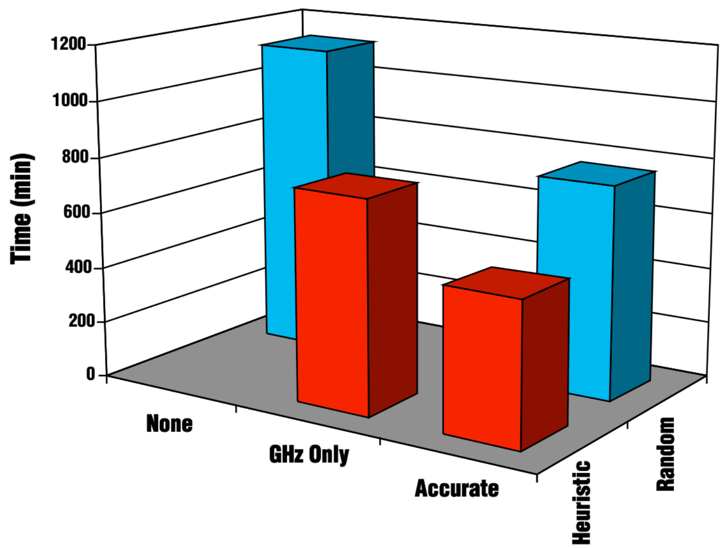
- A Virtual Grid (VG) takes
  - Shared heterogeneous resources
  - Scalable information service
- and provides
  - An hierarchy of application-defined aggregations (e.g. ClusterOf) with constraints (e.g. processor type) and rankings
- Virtual Grid Execution System (vgES) implements VG
  - VG Definition Language (vgDL)
  - VG Find And Bind (vgFAB)
  - VG Monitor (vgMON)
  - VG Application Launch (VgLAUNCH+DVCW)
  - VG Resource Info (vgAgent)



VGrADS Tool Research

- Scheduling of workflow computations
  - Off-line look-ahead scheduling dramatically improves in total time
  - Accurate performance models significantly affect quality of scheduling
  - Batch queue behavior can be predicted accurately enough for scheduling decisions
- Fault tolerance
  - Diskless checkpointing for linear algebra computations (application-specific)
  - Temporal reasoning for fault prediction
  - Optimal checkpoint frequency for iterative applications

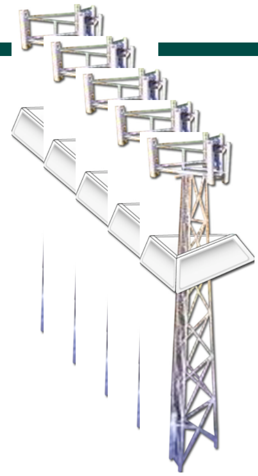




VGrADS: What's New

- SC'04
  - Scheduling EMAN application
    - Aware of performance models
- SC'05
  - Find and Bind (FAB) for resource selection
  - Scheduling EMAN application
    - Aware of batch queue predictions (and performance models)
- SC'06
  - Virtual Grid "slots" for resource availability
    - Start time + duration
    - Uses advance reservations where available
    - Uses batch queue prediction elsewhere
  - Scheduling LEAD application
    - Aware of reservations and batch queue predictions (and performance models)

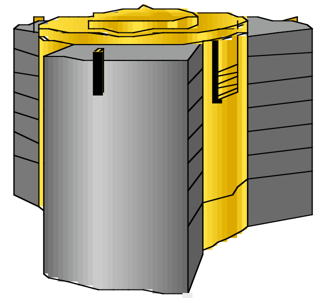
The LEAD Vision: A Paradigm Shift



DYNAMIC OBSERVATIONS

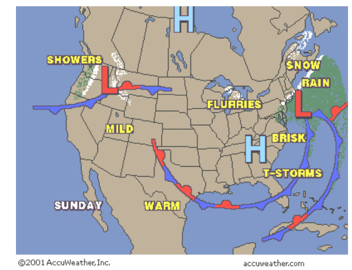
Analysis/Assimilation

Quality Control
Retrieval of Unobserved Quantities
Creation of Gridded Fields



Prediction/Detection

PCs to Teraflop Systems



Product Generation, Display, Dissemination

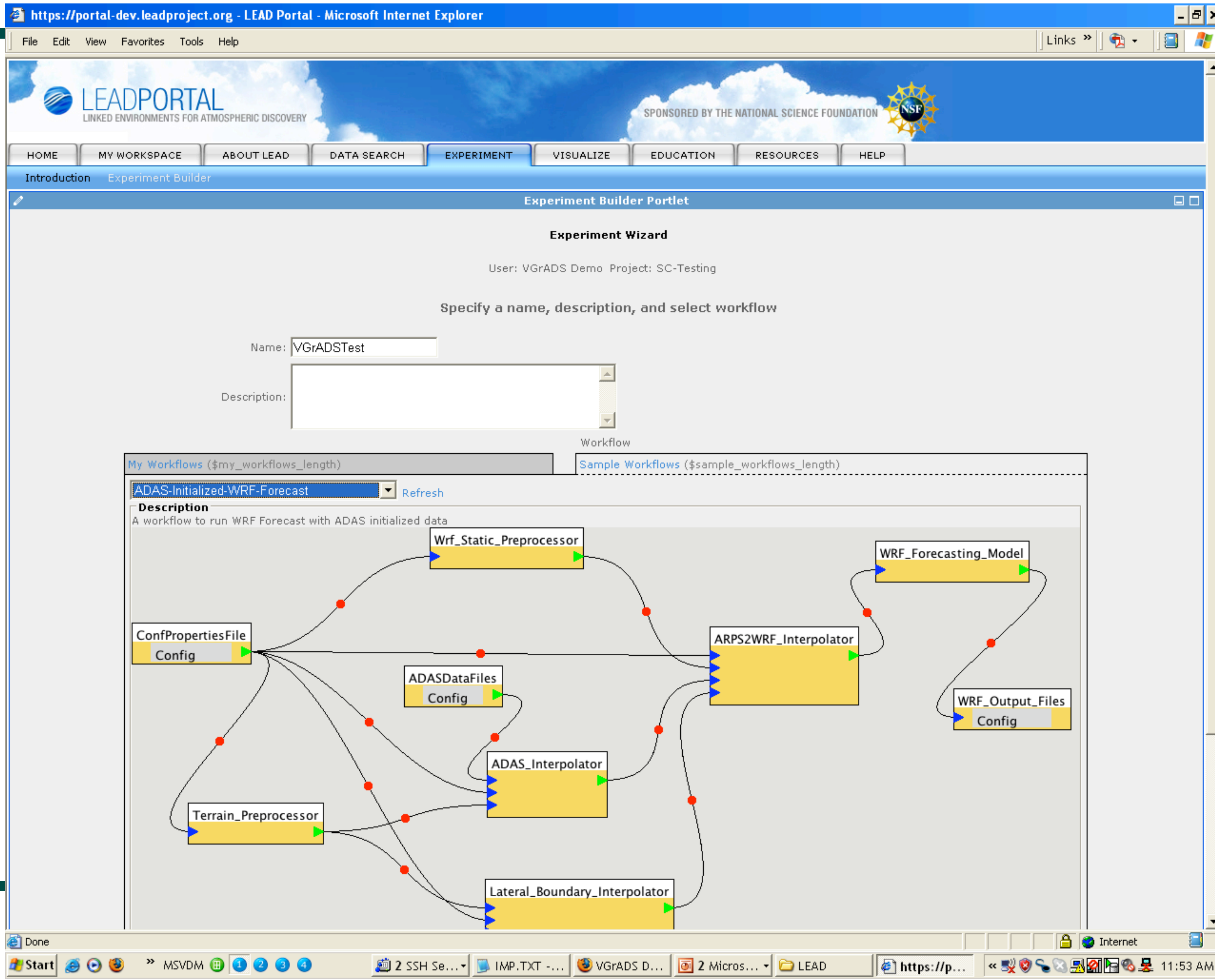
End Users

NWS
Private Companies
Students

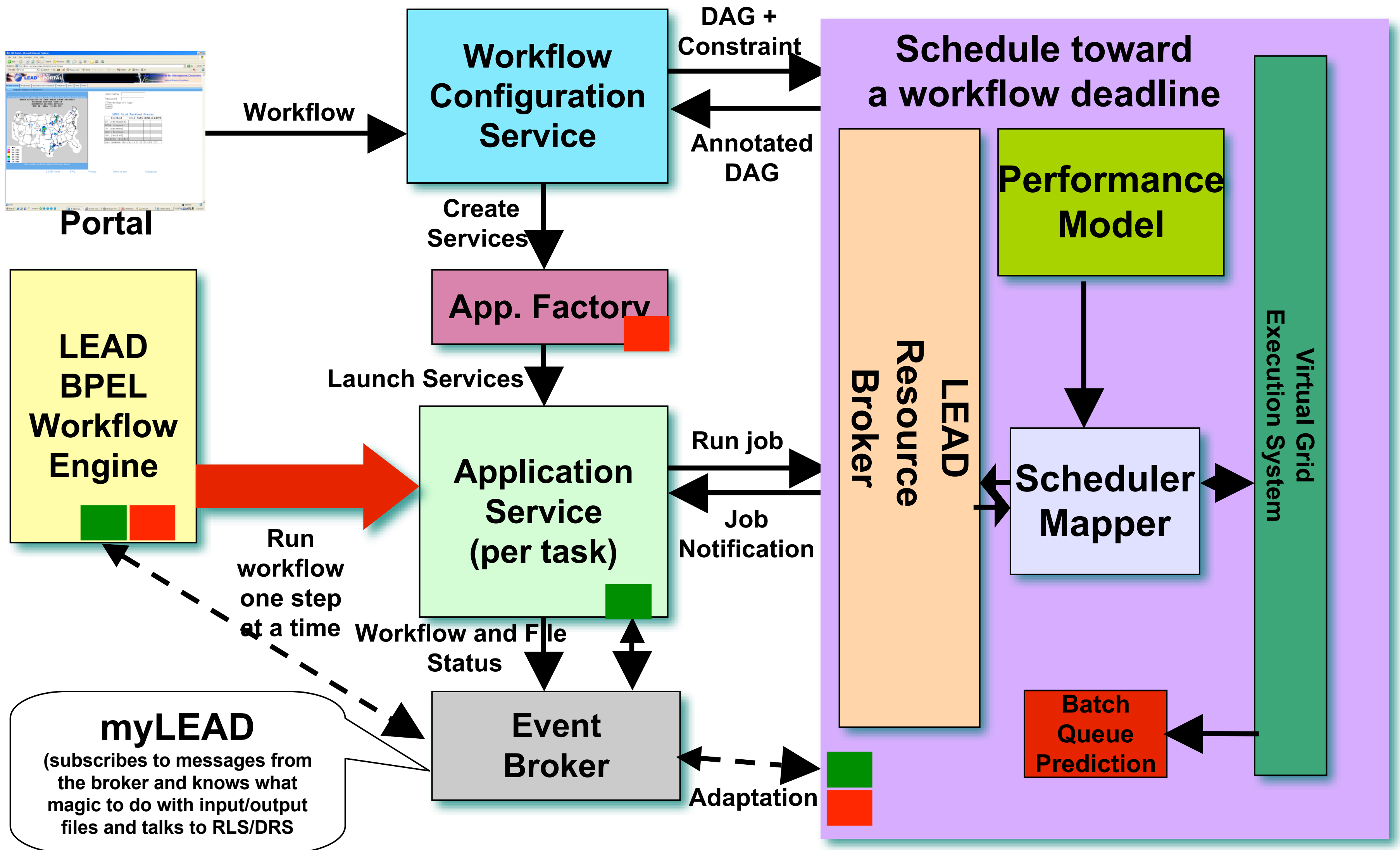
Models and Algorithms Driving Sensors

The CS challenge: Build cyberinfrastructure services that provide adaptability, scalability, availability, useability, and real-time response.

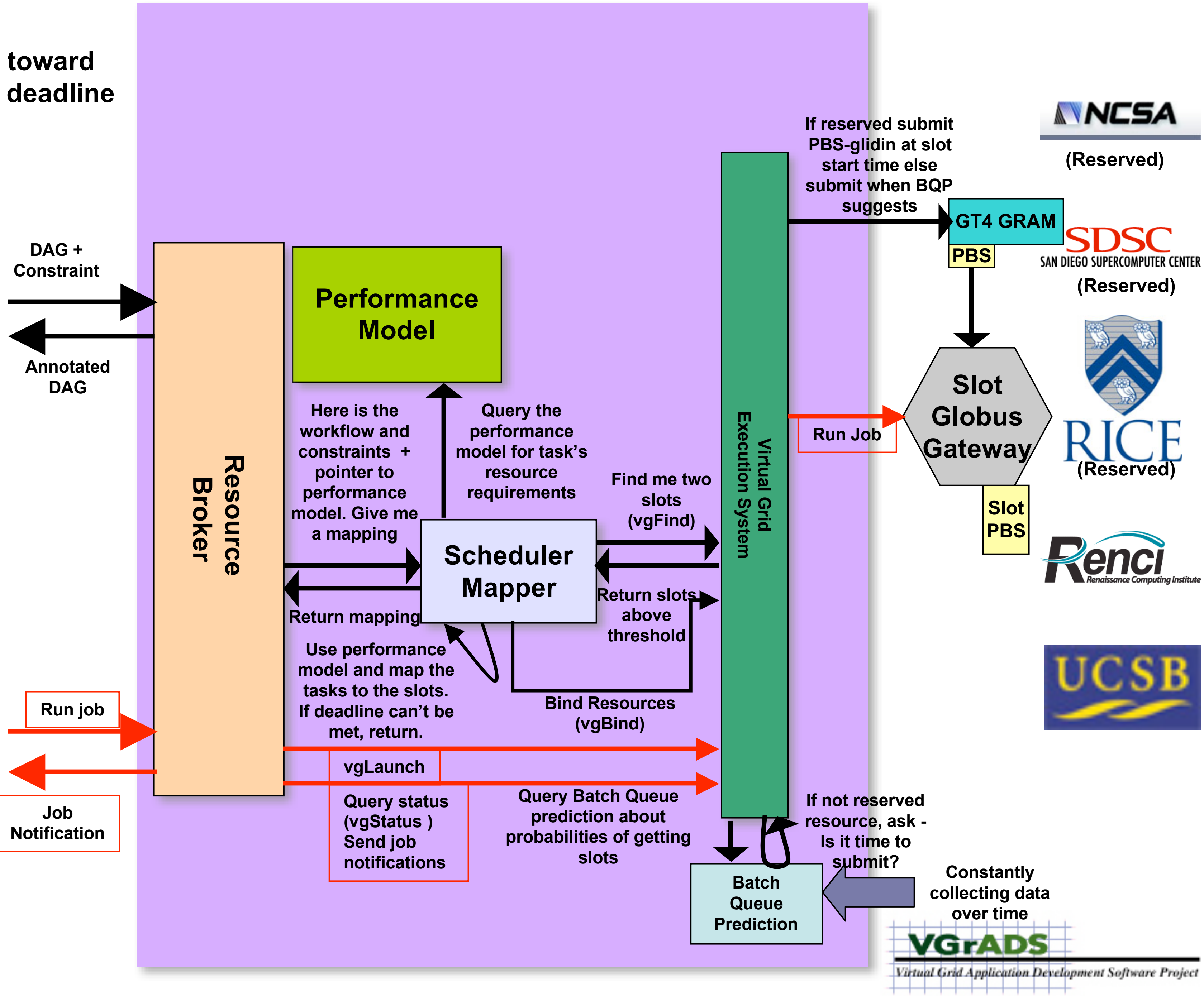
LEAD Portal – Experiment Builder



VGrADS Application Collaboration



Schedule toward a workflow deadline



Some Future Challenges

- Parallelism in the LEAD workflow manager
  - Parallel steps in different slots or within one slot
- Accurate Slot Requests Through Preliminary Scheduling
  - Minimization of wasted slot time
    - Accurate scheduling, better queue prediction
    - Dynamic adaptation of slot reservations
  - Requires some form of resource equivalence:
    - For step B, I need the equivalent of 200 Opterons, where 1 Opteron = 3 Itanium = 1.3 Power 5 (from perf models)
- Increased Schedule Robustness
  - Minimizing variation along the critical path
- Scheduling to Minimize Cost
  - In the presence of cycle exchange rates
  - Get the minimum-cost resources to solve the problem by the given deadline

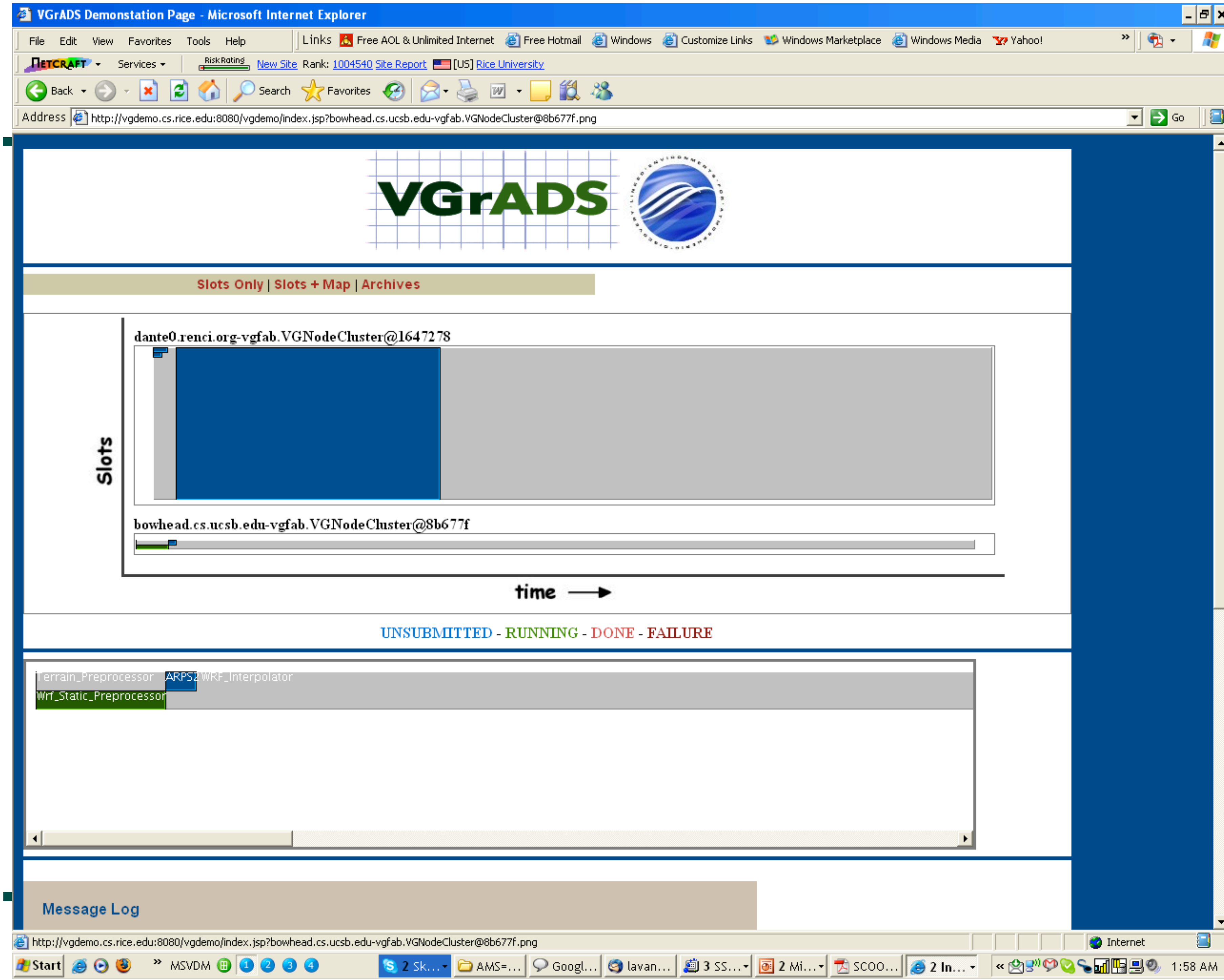
VGrADS at SC'06

- Booth Talks and Demos
  - Tuesday, noon - GCAS booth (1825)
  - Tuesday, 2:30 - USC booth (2246) [Not live]
  - Wednesday, 1:00 - SDSC booth (1915)
  - Thursday, 10:30 - RENCI booth (1143)
  - What you'll see
    - LEAD running on several clusters
    - Scheduler mapping LEAD components to slots
    - vgES managing slots via batch queue prediction
- Papers
  - "Improving Grid Resource Allocation via Integrated Selection and Binding" by Kee, et al. - Wednesday, 10:30
  - "Toward a Doctrine of Containtment: Grid Hosting with Adaptive Resource Control" by Ramakrishnan, et al. - Wednesday, 11:00
  - "Evaluation of a Workflow Scheduler Using Integrated Performance Modeling and Batch Queue Wait Time Prediction" by Nurmi, et al. - Thursday, 2:00

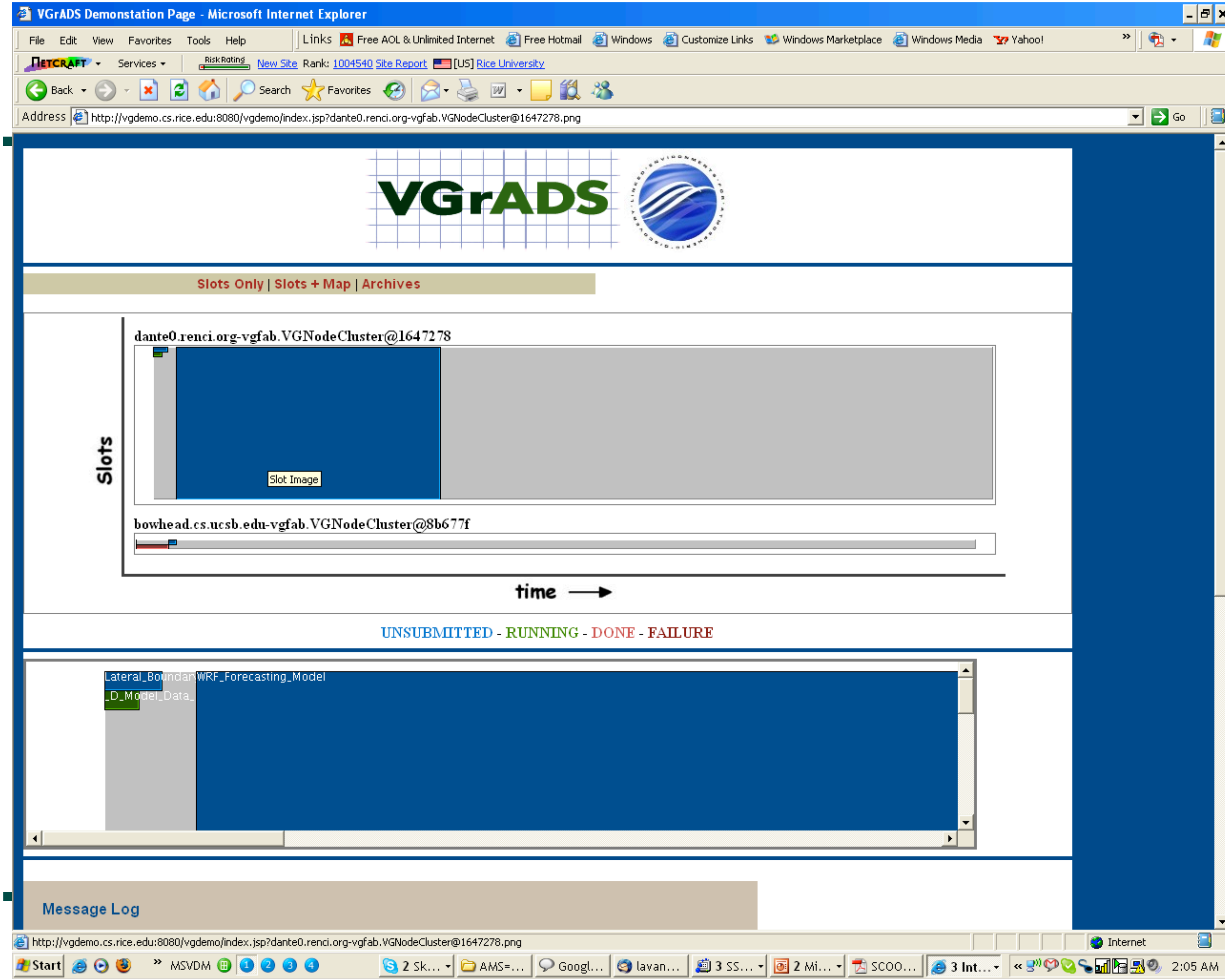
VGrADS
Virtual Grid Application Development Software Project

Launching from the LEAD Portal

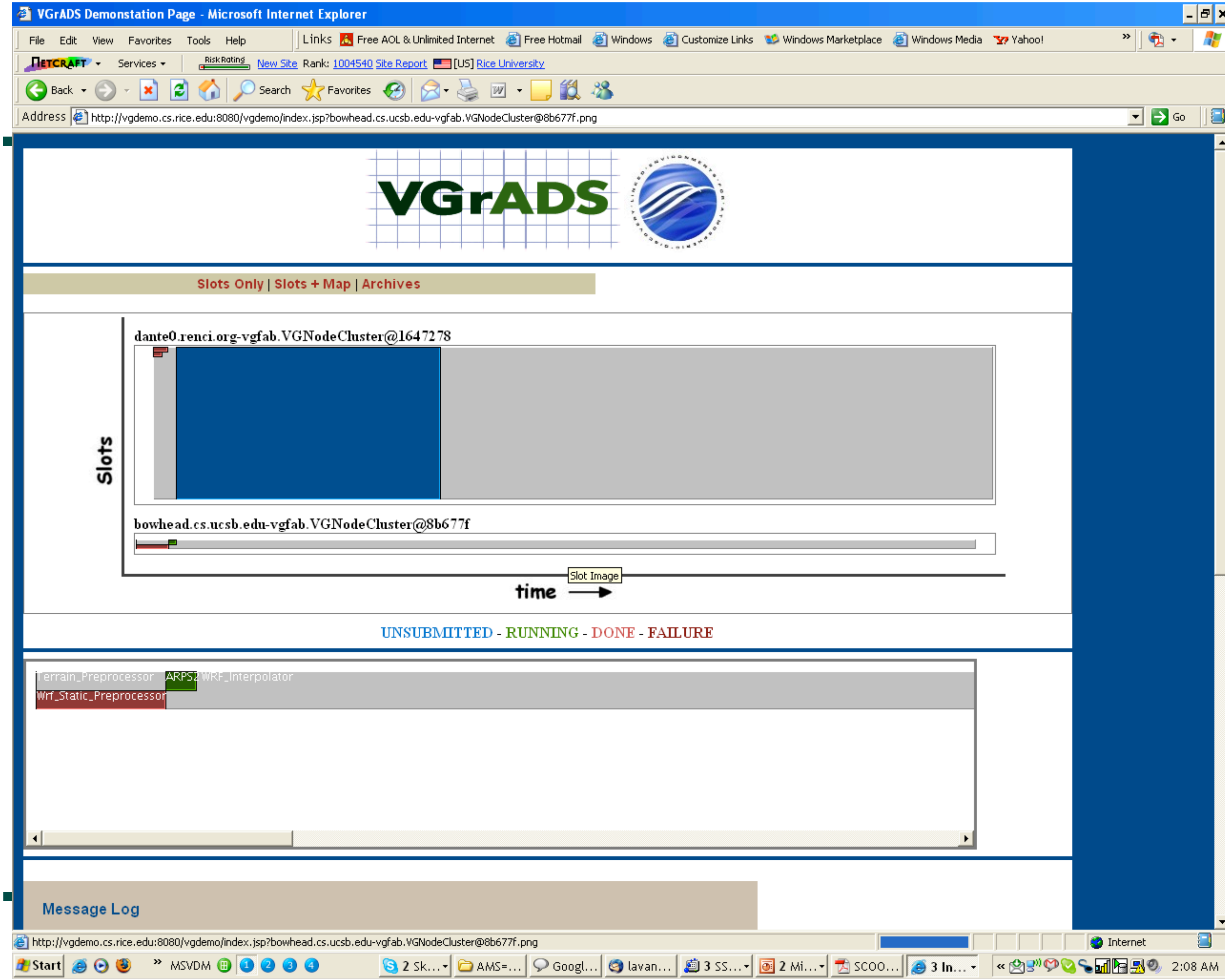
- **Work in Progress**


VGrADS Demonstation Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Free AOL & Unlimited Internet Free Hotmail Windows Customize Links Windows Marketplace Windows Media Yahoo!
Services
New Site Rank: 1004540 Site Report [US] Rice University
Back Search Favorites
Address http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f.png
Go
VGrADS
Slots Only | Slots + Map | Archives
dante0.renci.org-vgfab.VGNodeCluster@1647278
Slots
bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f
time
UNSUBMITTED - RUNNING - DONE - FAILURE
Terrain_Preprocessor
ARPS2
WRF_Interpolator
Wrf_Static_Preprocessor
Message Log
http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f.png
Internet
Start
MSVDM
1:58 AM

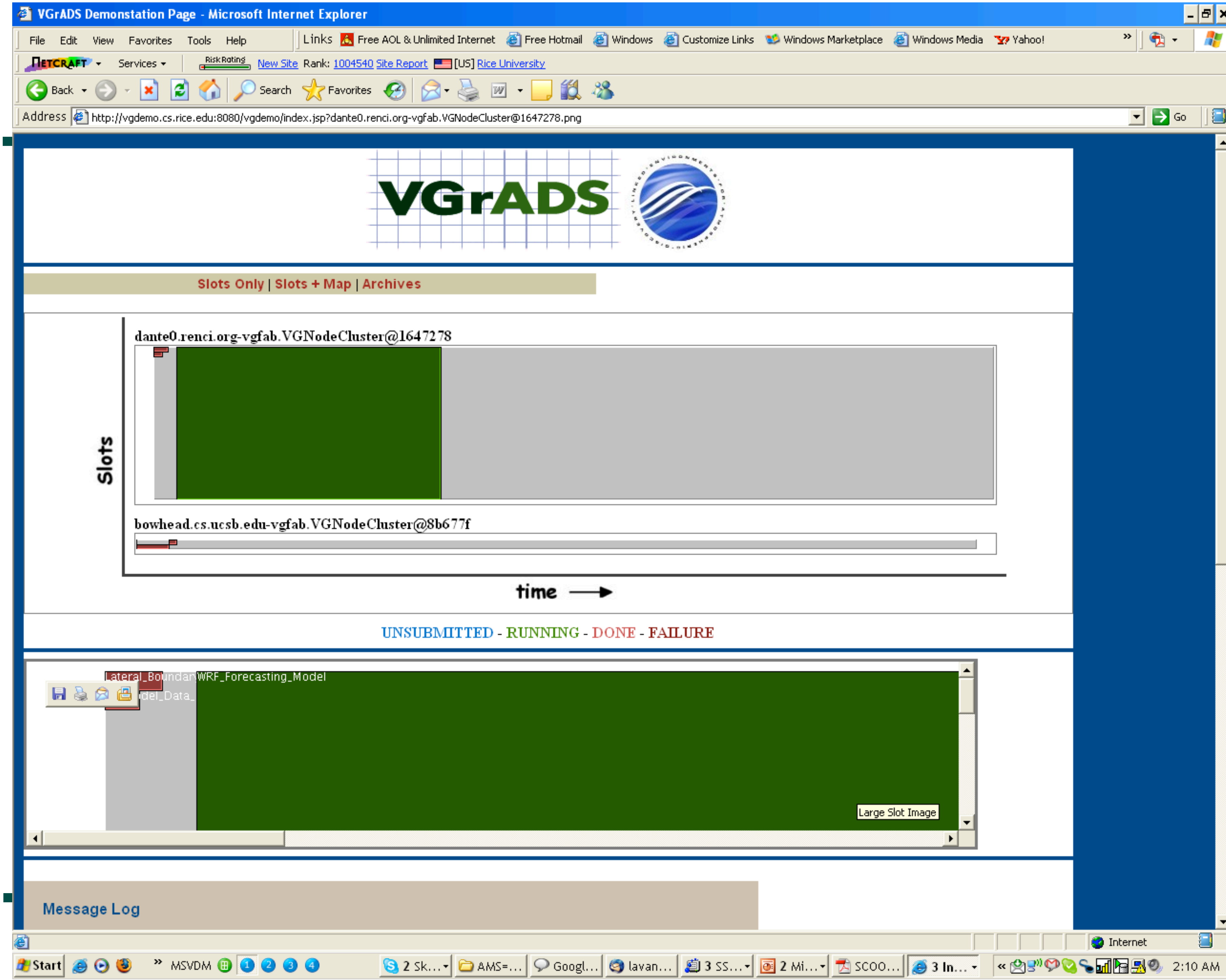
Project


VGrADS Demonstation Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Free AOL & Unlimited Internet Free Hotmail Windows Customize Links Windows Marketplace Windows Media Yahoo!
Services Risk Rating New Site Rank: 1004540 Site Report [US] Rice University
Back Search Favorites
Address http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?dante0.renci.org-vgfab.VGNodeCluster@1647278.png
Go
VGrADS
Slots Only | Slots + Map | Archives
Slots
dante0.renci.org-vgfab.VGNodeCluster@1647278
Slot Image
bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f
time
UNSUBMITTED - RUNNING - DONE - FAILURE
Lateral_Boundar
WRF_Forecasting_Model
_D_Model_Data_
Message Log
http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?dante0.renci.org-vgfab.VGNodeCluster@1647278.png
Internet
Start
MSVDM
2:05 AM

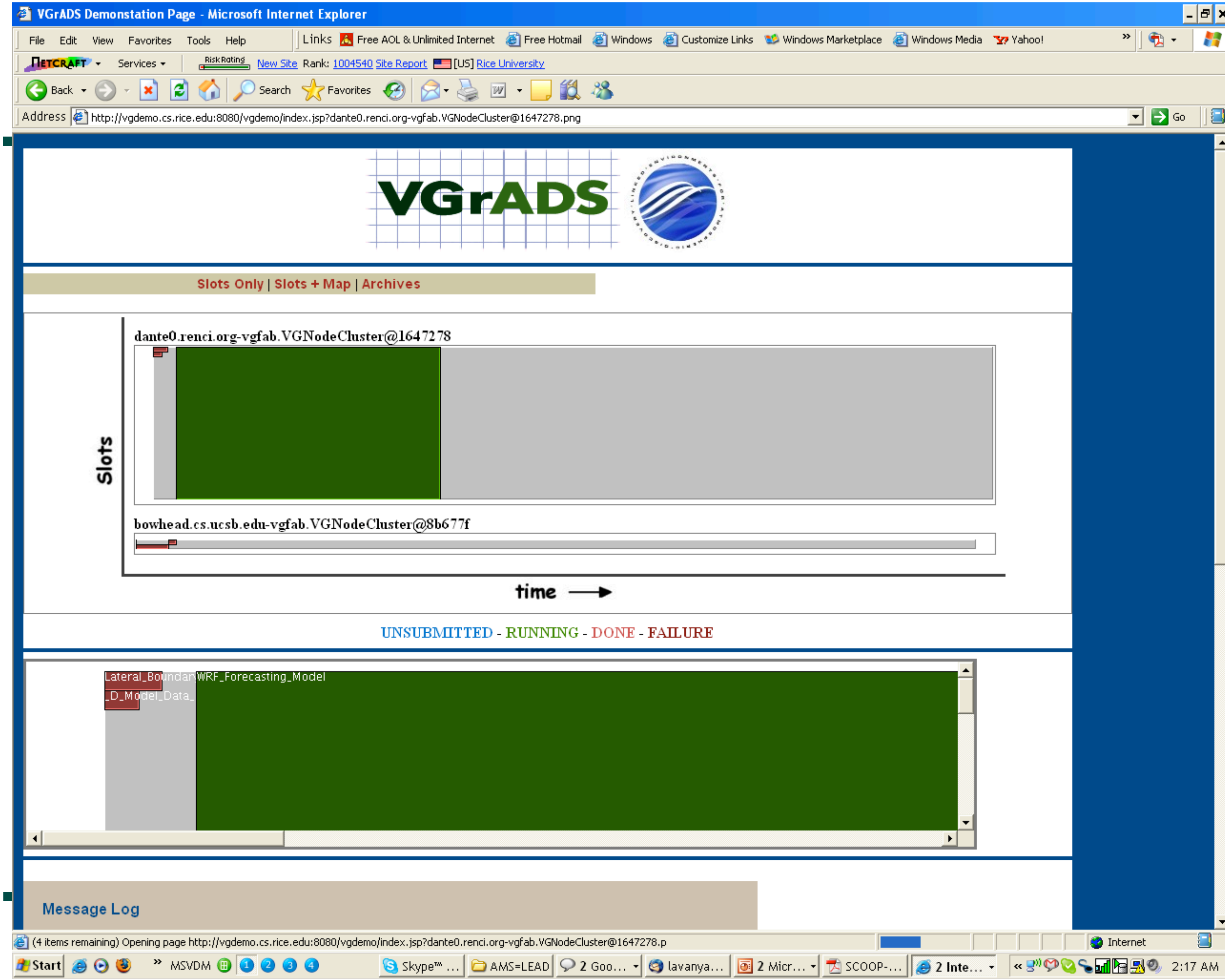
Project


VGrADS Demonstation Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Free AOL & Unlimited Internet Free Hotmail Windows Customize Links Windows Marketplace Windows Media Yahoo!
Services
New Site Rank: 1004540 Site Report [US] Rice University
Back Search Favorites
Address http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f.png
Go
VGrADS
Slots Only | Slots + Map | Archives
Slots
dante0.renci.org-vgfab.VGNodeCluster@1647278
bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f
Slot Image
time
UNSUBMITTED - RUNNING - DONE - FAILURE
Terrain_Preprocessor
ARPS2
WRF_Interpolator
Wrf_Static_Preprocessor
Message Log
http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f.png
Internet
Start
MSVDM
2:08 AM

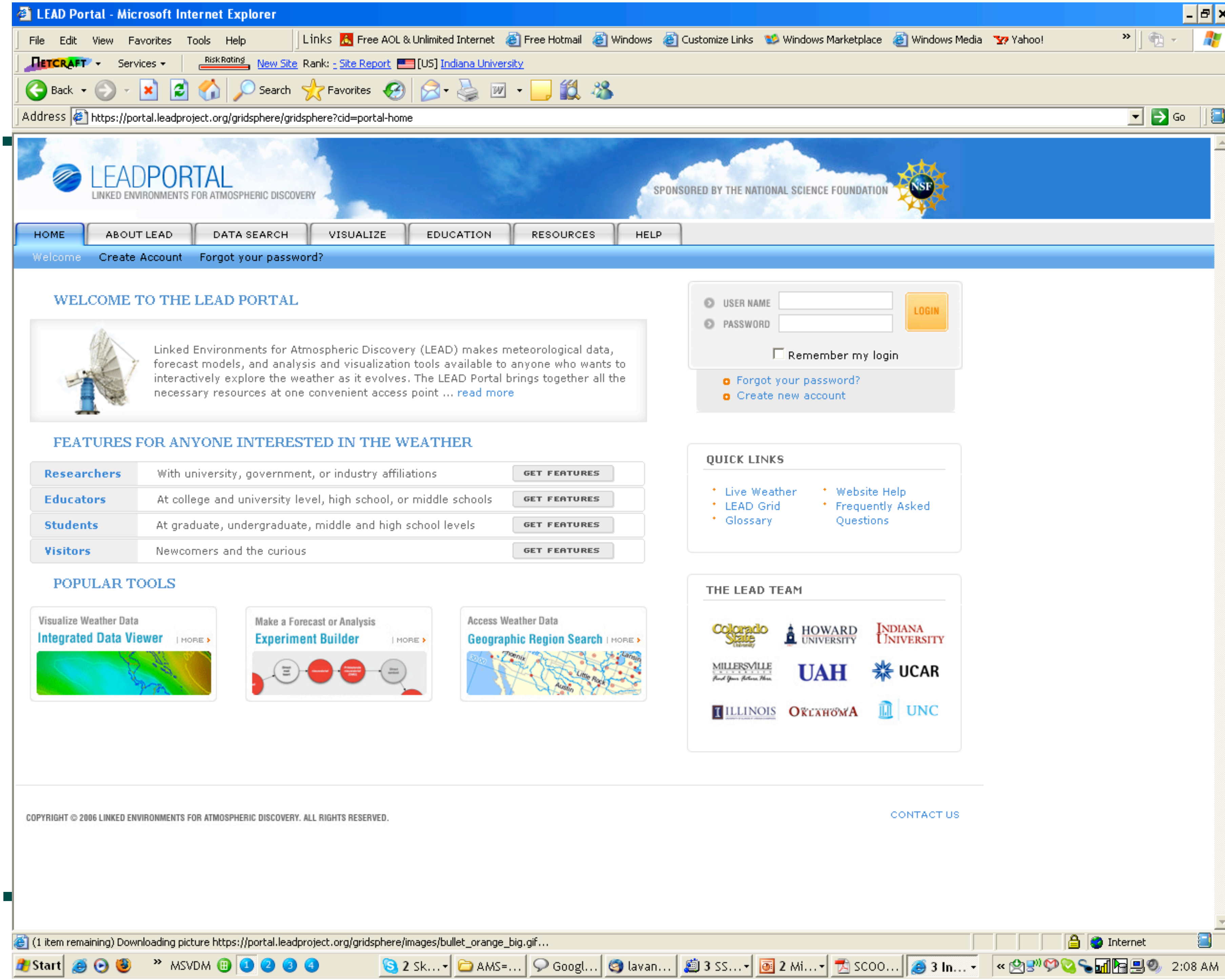
Project


VGrADS Demonstation Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Free AOL & Unlimited Internet Free Hotmail Windows Customize Links Windows Marketplace Windows Media Yahoo!
NETCRAFT Services Risk Rating New Site Rank: 1004540 Site Report [US] Rice University
Back Search Favorites
Address http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?dante0.renci.org-vgfab.VGNodeCluster@1647278.png
Go
VGrADS
Slots Only | Slots + Map | Archives
dante0.renci.org-vgfab.VGNodeCluster@1647278
Slots
bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f
time
UNSUBMITTED - RUNNING - DONE - FAILURE
WRF_Forecasting_Model
Large Slot Image
Message Log
Internet
Start
MSVDM
2:10 AM

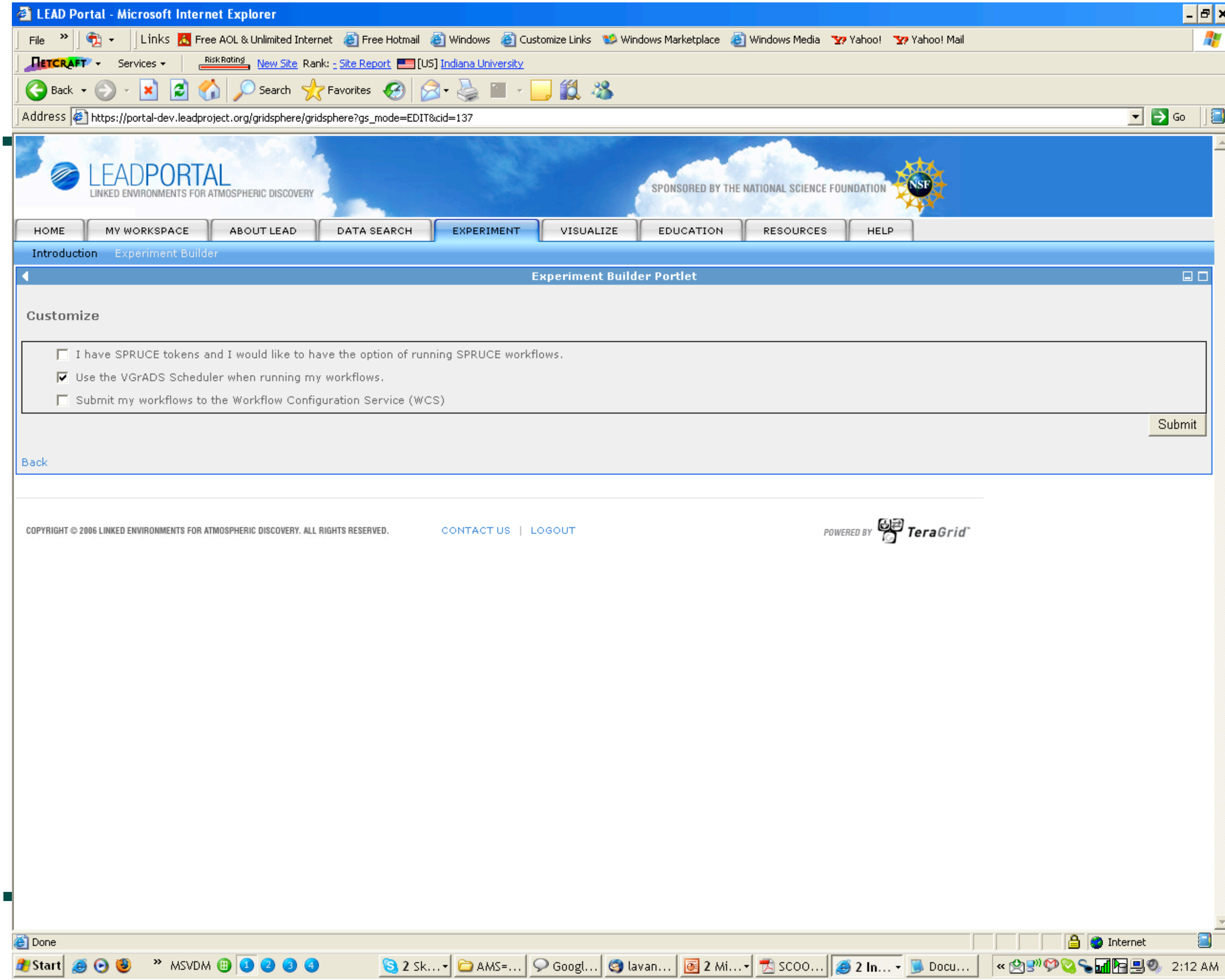
Project


VGrADS Demonstation Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Free AOL & Unlimited Internet Free Hotmail Windows Customize Links Windows Marketplace Windows Media Yahoo!
Services
New Site Rank: 1004540 Site Report [US] Rice University
Back Search Favorites
Address http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?dante0.renci.org-vgfab.VGNodeCluster@1647278.png
Go
VGrADS
Slots Only | Slots + Map | Archives
dante0.renci.org-vgfab.VGNodeCluster@1647278
Slots
bowhead.cs.ucsb.edu-vgfab.VGNodeCluster@8b677f
time
UNSUBMITTED - RUNNING - DONE - FAILURE
Lateral_Boundar
_D_Model_Data_
WRF_Forecasting_Model
Message Log
(4 items remaining) Opening page http://vgdemo.cs.rice.edu:8080/vgdemo/index.jsp?dante0.renci.org-vgfab.VGNodeCluster@1647278.p
Internet
Start
MSVDM
Skype™ ...
AMS=LEAD
2 Goo...
lavanya...
2 Micr...
SCOOP-...
2 Inte...
2:17 AM

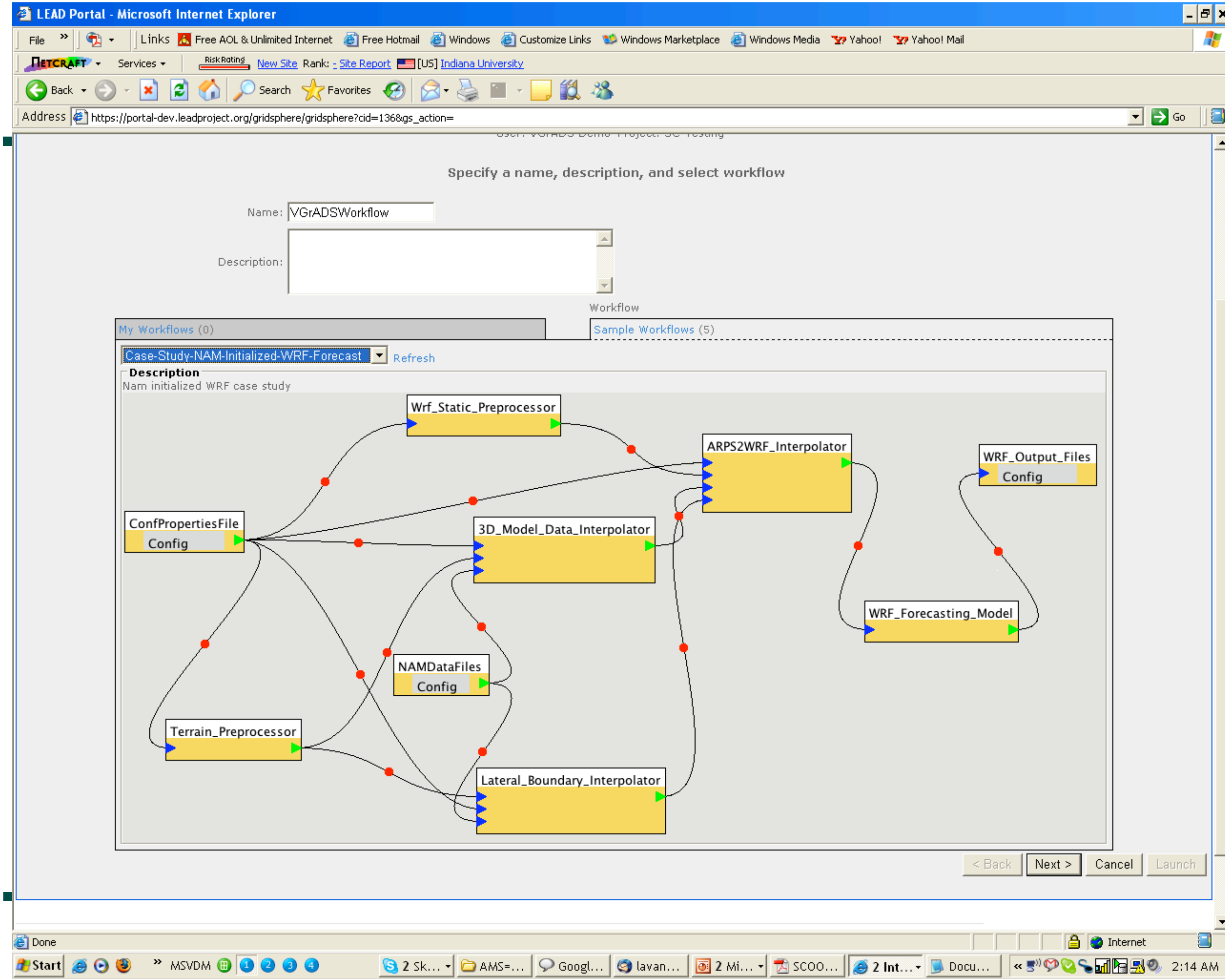
Project


LEAD Portal - Microsoft Internet Explorer
https://portal.leadproject.org/gridsphere/gridsphere?cid=portal-home
LEADPORTAL
LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY
SPONSORED BY THE NATIONAL SCIENCE FOUNDATION
HOME
ABOUT LEAD
DATA SEARCH
VISUALIZE
EDUCATION
RESOURCES
HELP
Welcome
Create Account
Forgot your password?
WELCOME TO THE LEAD PORTAL
Linked Environments for Atmospheric Discovery (LEAD) makes meteorological data, forecast models, and analysis and visualization tools available to anyone who wants to interactively explore the weather as it evolves. The LEAD Portal brings together all the necessary resources at one convenient access point ... read more
FEATURES FOR ANYONE INTERESTED IN THE WEATHER
Researchers
With university, government, or industry affiliations
GET FEATURES
Educators
At college and university level, high school, or middle schools
GET FEATURES
Students
At graduate, undergraduate, middle and high school levels
GET FEATURES
Visitors
Newcomers and the curious
GET FEATURES
POPULAR TOOLS
Visualize Weather Data
Integrated Data Viewer
Make a Forecast or Analysis
Experiment Builder
Access Weather Data
Geographic Region Search
USER NAME
PASSWORD
LOGIN
Remember my login
Forgot your password?
Create new account
QUICK LINKS
Live Weather
LEAD Grid
Glossary
Website Help
Frequently Asked Questions
THE LEAD TEAM
Colorado State University
HOWARD UNIVERSITY
INDIANA UNIVERSITY
MILLERSVILLE
UAH
UCAR
ILLINOIS
OKLAHOMA
UNC
COPYRIGHT © 2006 LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY. ALL RIGHTS RESERVED.
CONTACT US

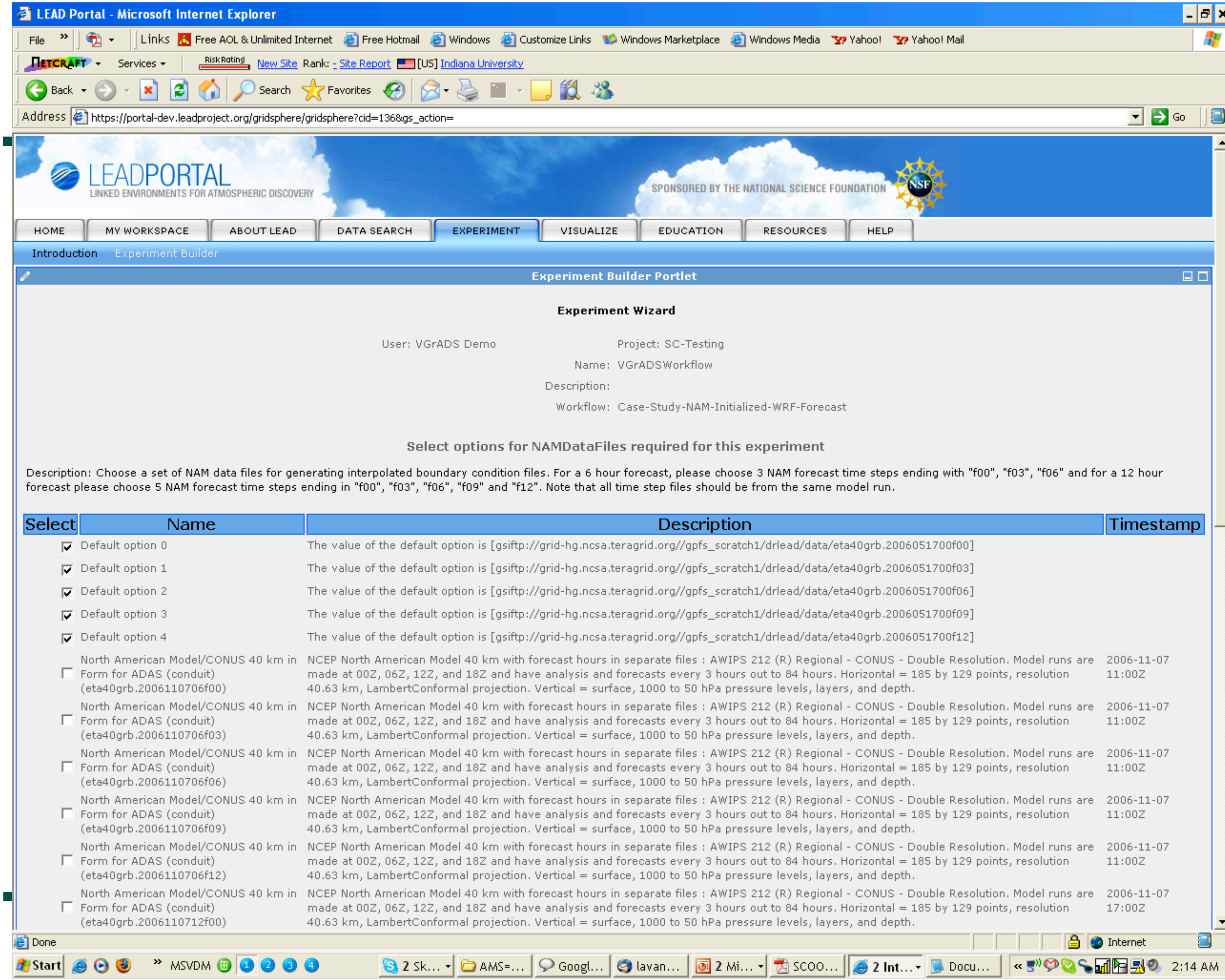
Project


LEAD Portal - Microsoft Internet Explorer
Address https://portal-dev.leadproject.org/gridsphere/gridsphere?gs_mode=EDIT&cid=137
LEADPORTAL
LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY
SPONSORED BY THE NATIONAL SCIENCE FOUNDATION
HOME MY WORKSPACE ABOUT LEAD DATA SEARCH EXPERIMENT VISUALIZE EDUCATION RESOURCES HELP
Introduction Experiment Builder
Experiment Builder Portlet
Customize
I have SPRUCE tokens and I would like to have the option of running SPRUCE workflows.
Use the VGrADS Scheduler when running my workflows.
Submit my workflows to the Workflow Configuration Service (WCS)
Submit
Back
COPYRIGHT © 2006 LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY. ALL RIGHTS RESERVED.
CONTACT US | LOGOUT
POWERED BY TeraGrid
Done
Internet
2:12 AM

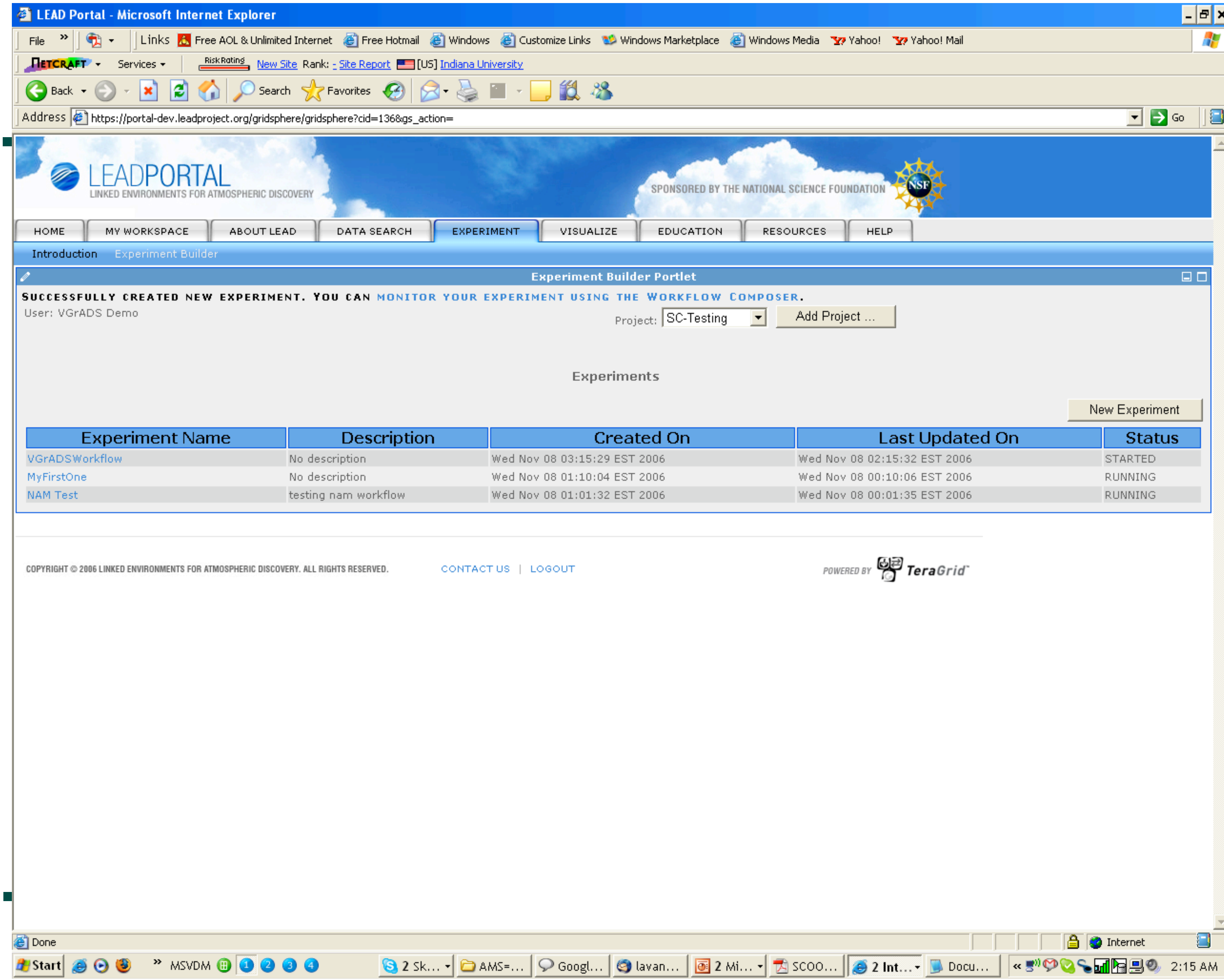
Project


LEAD Portal - Microsoft Internet Explorer
File
Links
Free AOL & Unlimited Internet
Free Hotmail
Windows
Customize Links
Windows Marketplace
Windows Media
Yahoo!
Yahoo! Mail
Services
Risk Rating
New Site
Rank: - Site Report
[US] Indiana University
Back
Search
Favorites
Address
https://portal-dev.leadproject.org/gridsphere/gridsphere?cid=136&gs_action=
Go
Specify a name, description, and select workflow
Name:
VGrADSWorkflow
Description:
Workflow
My Workflows (0)
Sample Workflows (5)
Case-Study-NAM-Initialized-WRF-Forecast
Refresh
Description
Nam initialized WRF case study
Wrf_Static_Preprocessor
ARPS2WRF_Interpolator
WRF_Output_Files
Config
ConfPropertiesFile
Config
3D_Model_Data_Interpolator
WRF_Forecasting_Model
NAMDataFiles
Config
Terrain_Preprocessor
Lateral_Boundary_Interpolator
< Back
Next >
Cancel
Launch
Done
Internet
Start
MSVDM
2:14 AM

Project


LEAD Portal - Microsoft Internet Explorer
File
Links Free AOL & Unlimited Internet Free Hotmail Windows Customize Links Windows Marketplace Windows Media Yahoo! Yahoo! Mail
NETCRAFT Services Risk Rating New Site Rank: - Site Report [US] Indiana University
Back Search Favorites
Address https://portal-dev.leadproject.org/gridsphere/gridsphere?cid=136&gs_action=
Go
LEADPORTAL
LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY
SPONSORED BY THE NATIONAL SCIENCE FOUNDATION
NSF
HOME
MY WORKSPACE
ABOUT LEAD
DATA SEARCH
EXPERIMENT
VISUALIZE
EDUCATION
RESOURCES
HELP
Introduction
Experiment Builder
Experiment Builder Portlet
Experiment Wizard
User: VGrADS Demo
Project: SC-Testing
Name: VGrADSWorkflow
Description:
Workflow: Case-Study-NAM-Initialized-WRF-Forecast
Select options for NAMDataFiles required for this experiment
Description: Choose a set of NAM data files for generating interpolated boundary condition files. For a 6 hour forecast, please choose 3 NAM forecast time steps ending with "f00", "f03", "f06" and for a 12 hour forecast please choose 5 NAM forecast time steps ending in "f00", "f03", "f06", "f09" and "f12". Note that all time step files should be from the same model run.
Select
Name
Description
Timestamp
Default option 0
The value of the default option is [gsiftp://grid-hg.ncsa.teragrid.org//gpfs_scratch1/drlead/data/eta40grb.2006051700f00]
Default option 1
The value of the default option is [gsiftp://grid-hg.ncsa.teragrid.org//gpfs_scratch1/drlead/data/eta40grb.2006051700f03]
Default option 2
The value of the default option is [gsiftp://grid-hg.ncsa.teragrid.org//gpfs_scratch1/drlead/data/eta40grb.2006051700f06]
Default option 3
The value of the default option is [gsiftp://grid-hg.ncsa.teragrid.org//gpfs_scratch1/drlead/data/eta40grb.2006051700f09]
Default option 4
The value of the default option is [gsiftp://grid-hg.ncsa.teragrid.org//gpfs_scratch1/drlead/data/eta40grb.2006051700f12]
North American Model/CONUS 40 km in Form for ADAS (conduit) (eta40grb.2006110706f00)
NCEP North American Model 40 km with forecast hours in separate files : AWIPS 212 (R) Regional - CONUS - Double Resolution. Model runs are made at 00Z, 06Z, 12Z, and 18Z and have analysis and forecasts every 3 hours out to 84 hours. Horizontal = 185 by 129 points, resolution 40.63 km, LambertConformal projection. Vertical = surface, 1000 to 50 hPa pressure levels, layers, and depth.
2006-11-07 11:00Z
North American Model/CONUS 40 km in Form for ADAS (conduit) (eta40grb.2006110706f03)
NCEP North American Model 40 km with forecast hours in separate files : AWIPS 212 (R) Regional - CONUS - Double Resolution. Model runs are made at 00Z, 06Z, 12Z, and 18Z and have analysis and forecasts every 3 hours out to 84 hours. Horizontal = 185 by 129 points, resolution 40.63 km, LambertConformal projection. Vertical = surface, 1000 to 50 hPa pressure levels, layers, and depth.
2006-11-07 11:00Z
North American Model/CONUS 40 km in Form for ADAS (conduit) (eta40grb.2006110706f06)
NCEP North American Model 40 km with forecast hours in separate files : AWIPS 212 (R) Regional - CONUS - Double Resolution. Model runs are made at 00Z, 06Z, 12Z, and 18Z and have analysis and forecasts every 3 hours out to 84 hours. Horizontal = 185 by 129 points, resolution 40.63 km, LambertConformal projection. Vertical = surface, 1000 to 50 hPa pressure levels, layers, and depth.
2006-11-07 11:00Z
North American Model/CONUS 40 km in Form for ADAS (conduit) (eta40grb.2006110706f09)
NCEP North American Model 40 km with forecast hours in separate files : AWIPS 212 (R) Regional - CONUS - Double Resolution. Model runs are made at 00Z, 06Z, 12Z, and 18Z and have analysis and forecasts every 3 hours out to 84 hours. Horizontal = 185 by 129 points, resolution 40.63 km, LambertConformal projection. Vertical = surface, 1000 to 50 hPa pressure levels, layers, and depth.
2006-11-07 11:00Z
North American Model/CONUS 40 km in Form for ADAS (conduit) (eta40grb.2006110706f12)
NCEP North American Model 40 km with forecast hours in separate files : AWIPS 212 (R) Regional - CONUS - Double Resolution. Model runs are made at 00Z, 06Z, 12Z, and 18Z and have analysis and forecasts every 3 hours out to 84 hours. Horizontal = 185 by 129 points, resolution 40.63 km, LambertConformal projection. Vertical = surface, 1000 to 50 hPa pressure levels, layers, and depth.
2006-11-07 11:00Z
North American Model/CONUS 40 km in Form for ADAS (conduit) (eta40grb.2006110712f00)
NCEP North American Model 40 km with forecast hours in separate files : AWIPS 212 (R) Regional - CONUS - Double Resolution. Model runs are made at 00Z, 06Z, 12Z, and 18Z and have analysis and forecasts every 3 hours out to 84 hours. Horizontal = 185 by 129 points, resolution 40.63 km, LambertConformal projection. Vertical = surface, 1000 to 50 hPa pressure levels, layers, and depth.
2006-11-07 17:00Z
Done
Internet
Start
MSVDM
2 Sk...
AMS=...
Googl...
lavan...
2 Mi...
SCOO...
2 Int...
Docu...
2:14 AM

Project



LEAD Portal - Microsoft Internet Explorer
Address https://portal-dev.leadproject.org/gridsphere/gridsphere?cid=136&gs_action=
LEADPORTAL
LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY
SPONSORED BY THE NATIONAL SCIENCE FOUNDATION
HOME
MY WORKSPACE
ABOUT LEAD
DATA SEARCH
EXPERIMENT
VISUALIZE
EDUCATION
RESOURCES
HELP
Introduction
Experiment Builder
Experiment Builder Portlet
SUCCESSFULLY CREATED NEW EXPERIMENT. YOU CAN MONITOR YOUR EXPERIMENT USING THE WORKFLOW COMPOSER.
User: VGrADS Demo
Project: SC-Testing
Add Project ...
Experiments
New Experiment
Experiment Name | Description | Created On | Last Updated On | Status
VGrADSWorkflow | No description | Wed Nov 08 03:15:29 EST 2006 | Wed Nov 08 02:15:32 EST 2006 | STARTED
MyFirstOne | No description | Wed Nov 08 01:10:04 EST 2006 | Wed Nov 08 00:10:06 EST 2006 | RUNNING
NAM Test | testing nam workflow | Wed Nov 08 01:01:32 EST 2006 | Wed Nov 08 00:01:35 EST 2006 | RUNNING
COPYRIGHT © 2006 LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY. ALL RIGHTS RESERVED.
CONTACT US | LOGOUT
POWERED BY TeraGrid
Done
Internet
Start
2:15 AM

Project

Scheduling with Batch Queues

- Last Year: VGrADS supported scheduling using estimated batch queue waiting times
  - Batch queue estimates are factored into communication time
    - E.g., the delay in moving from one resource to another is data movement time + estimated batch queue waiting time
  - Unfortunately, estimates can have large standard deviations
- This Year: limiting variability through two strategies:
  - Resource reservations: partially supported on the TeraGrid and other schedulers
  - In advance queue insertion: submit jobs before data arrives based on estimates
    - Can be used to simulate advance reservations
- Exploiting this requires a preliminary schedule indicating when the resources are needed
  - Problem: how to build an accurate schedule when exact resource types are unknown

Preliminary Scheduling Solution

- Use performance models to specify alternative resources
  - For step B, I need the equivalent of 200 Opterons, where 1 Opteron = 3 Itanium = 1.3 Power 5
    - Equivalence from performance model
- This permits an accurate preliminary schedule because the performance model standardizes the time for each step
  - Scheduling can then proceed with accurate estimates of when each resource collection will be needed
  - Makes advance reservations more accurate
    - Data will arrive neither too early or too late
- It may provide a mixture to meet the computational requirements, if the specification permits
  - Give me a loose bag of tight bags containing the equivalent of 200 Opterons, minimize the number of tight bags and the overall cost
    - Solution might be 150 Opterons in one cluster and 150 Itaniums in another