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

Keith Cooper
Ken Kennedy
Charles Koelbel
Richard Tapia
Linda Torczon

Dan Reed

Jack Dongarra

Fran Berman
Andrew Chien
Henri Casanova

Rich Wolski

Carl Kesselman

Lennart Johnsson

• Plus many graduate students, postdocs, and technical staff!
Vision: Global Distributed Problem Solving

• Where We Want To Be
  o Transparent Grid computing
    - Submit job
    - Find & schedule resources
    - Execute efficiently

• Where We Are
  o Low-level hand programming
  o Programmer must manage:
    - Heterogeneous resources
    - Scheduling of computation and data movement
    - Fault tolerance and performance adaptation

• What Do We Propose as A Solution?
  o Separate application development from resource management
    - Through an abstraction called the Virtual Grid
  o 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
  o 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
  o Virtual Grid Execution System (vgES) produces specific virtual grid matching specification
  o Avoids need for scheduling against the entire space of global resources

• Generic In-Advance Scheduling of Application Workflows
  o Application includes performance models for all workflow nodes
    - Performance models automatically constructed
  o 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
  o Scheduling EMAN application
    - Aware of performance models

• SC’05
  o Find and Bind (FAB) for resource selection
  o Scheduling EMAN application
    - Aware of batch queue predictions (and performance models)

• SC’06
  o Virtual Grid "slots" for resource availability
    - Start time + duration
    - Uses advance reservations where available
    - Uses batch queue prediction elsewhere
  o Scheduling LEAD application
    - Aware of reservations and batch queue predictions (and performance models)
The LEAD Vision: A Paradigm Shift

The CS challenge: Build cyberinfrastructure services that provide adaptability, scalability, availability, useability, and real-time response.
LEAD Portal – Experiment Builder
VGrADS Application Collaboration

LEAD
Linked Environments for Atmospheric Discovery
Schedule toward a workflow deadline

- Dag + Constraint
- Annotated Dag
- DAG + Constraint

Resource Broker

- Performance Model
  - Here is the workflow and constraints + pointer to performance model. Give me a mapping
  - Query the performance model for task's resource requirements

Scheduler Mapper

- Return mapping
- Use performance model and map the tasks to the slots. If deadline can't be met, return.

Virtual Grid Execution System

- Batch Queue Prediction
  - Query Batch Queue prediction about probabilities of getting slots
- Bind Resources (vgBind)
  - Find me two slots (vgFind)
  - Return slots above threshold

Batch Queue Prediction

- GT4 Gram
- PBS
- Slot
  - SDS
  - Reserved
  - RICE
  - Reserved
  - UCSB

Slot Global Gateway

- GT4 Gram
- PBS
- Slot

PTSG

- (Reserved)
- (Reserved)
- (Reserved)

If reserved submit PBS-glidin at slot start time else submit when BQP suggests

Constantly collecting data over time

Run Job

- vglAunche
- Query status (vgStatus)
- Send job notifications

Job Notification

If reserved resource, ask - Is it time to submit?
Some Future Challenges

• Parallelism in the LEAD workflow manager
  o Parallel steps in different slots or within one slot

• Accurate Slot Requests Through Preliminary Scheduling
  o Minimization of wasted slot time
    - Accurate scheduling, better queue prediction
    - Dynamic adaptation of slot reservations
  o 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
  o Minimizing variation along the critical path

• Scheduling to Minimize Cost
  o In the presence of cycle exchange rates
  o Get the minimum-cost resources to solve the problem by the given deadline
VGrADS at SC’06

• Booth Talks and Demos
  o Tuesday, noon - GCAS booth (1825)
  o Tuesday, 2:30 - USC booth (2246) [Not live]
  o Wednesday, 1:00 - SDSC booth (1915)
  o Thursday, 10:30 - RENCI booth (1143)
  o What you’ll see
    - LEAD running on several clusters
    - Scheduler mapping LEAD components to slots
    - vgES managing slots via batch queue prediction

• Papers
  o “Improving Grid Resource Allocation via Integrated Selection and Binding” by Kee, et al. - Wednesday, 10:30
  o “Toward a Doctrine of Containment: Grid Hosting with Adaptive Resource Control” by Ramakrishnan, et al. - Wednesday, 11:00
Launching from the LEAD Portal

- Work in Progress
VGrADS

Slots Only | Slots + Map | Archives

Unsubmitted - Running - Done - Failure
VGrADS Demonstration Page - Microsoft Internet Explorer

UNSUBMITTED - RUNNING - DONE - FAILURE
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

Educators
At college and university level, high school, or middle schools

Students
At graduate, undergraduate, middle and high school levels

Visitors
Newcomers and the curious

QUICK LINKS

- Live Weather
- LEAD Grid
- Website Help
- Frequently Asked Questions

THE LEAD TEAM

Colorado State University
Howard University
Indiana University
Millennium State University
University of Alabama at Huntsville
University Corporation for Atmospheric Research
University of North Carolina

POPULAR TOOLS

Visualize Weather Data
Integrated Data Viewer

Make a Forecast or Analysis
Experiment Builder

Access Weather Data
Geographic Region Search
**Experiment Wizard**

**Project:** SC-Testing  
**Name:** VGRID Workflow  
**Description:**  
**Workflow:** Case-Study-NAM-Initialized-WRF-Forecast  

**Select options for NAMdataFiles required for this experiment**

<table>
<thead>
<tr>
<th>Select</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Default option 0</td>
<td>The value of the default option is [<a href="https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517000000">https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517000000</a>]</td>
</tr>
<tr>
<td></td>
<td>Default option 1</td>
<td>The value of the default option is [<a href="https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517010000">https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517010000</a>]</td>
</tr>
<tr>
<td></td>
<td>Default option 2</td>
<td>The value of the default option is [<a href="https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517020000">https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517020000</a>]</td>
</tr>
<tr>
<td></td>
<td>Default option 3</td>
<td>The value of the default option is [<a href="https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517030000">https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517030000</a>]</td>
</tr>
<tr>
<td></td>
<td>Default option 4</td>
<td>The value of the default option is [<a href="https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517040000">https://grid-hg.ncsa.tamu.edu/gridHOME/scratch/df/ele/eta4Grb.20060517040000</a>]</td>
</tr>
</tbody>
</table>

**Note:** 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 "630", "180", "450" and for a 12 hour forecast please choose 3 NAM forecast time steps ending in "630", "180", "450" and "540".
### Experiment Builder Portlet

Successfully created new experiment. You can monitor your experiment using the Workflow Composer.

User: WorldDemo

<table>
<thead>
<tr>
<th>Experiment Name</th>
<th>Description</th>
<th>Created On</th>
<th>Last Updated On</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGridADSWorkflow</td>
<td>No description</td>
<td>Wed Nov 08 00:38:29 EST 2005</td>
<td>Wed Nov 08 02:15:02 EST 2005</td>
<td>STARTED</td>
</tr>
<tr>
<td>MyFirstOne</td>
<td>No description</td>
<td>Wed Nov 08 01:10:04 EST 2005</td>
<td>Wed Nov 08 00:10:05 EST 2005</td>
<td>RUNNING</td>
</tr>
<tr>
<td>NAM Test</td>
<td>testing nam workflow</td>
<td>Wed Nov 08 01:10:56 EST 2005</td>
<td>Wed Nov 08 00:10:03 EST 2005</td>
<td>RUNNING</td>
</tr>
</tbody>
</table>

Experiments

Copyright © 2005 Linked Environments for Atmospheric Discovery, All Rights Reserved.
Scheduling with Batch Queues

• Last Year: VGrADS supported scheduling using estimated batch queue waiting times
  o 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
  o Unfortunately, estimates can have large standard deviations

• This Year: limiting variability through two strategies:
  o Resource reservations: partially supported on the TeraGrid and other schedulers
  o 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
  o Problem: how to build an accurate schedule when exact resource types are unknown
Preliminary Scheduling Solution

• Use performance models to specify alternative resources
  o 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
  o Scheduling can then proceed with accurate estimates of when each resource collection will be needed
  o 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
  o 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