## Fault Tolerance and Recovery for Grid Workflow Systems

Anirban Mandal Gopi Kandaswamy Dan Reed

VGrADS Workshop April 2007





#### **Presentation Outline**

- Motivation and rationale
- Fault tolerance and recovery service
- Fault tolerance and recovery algorithms
  - over-provisioning, simple restart, migration
- Proposed next steps
- Demonstration
  - LEAD-VGrADS integration







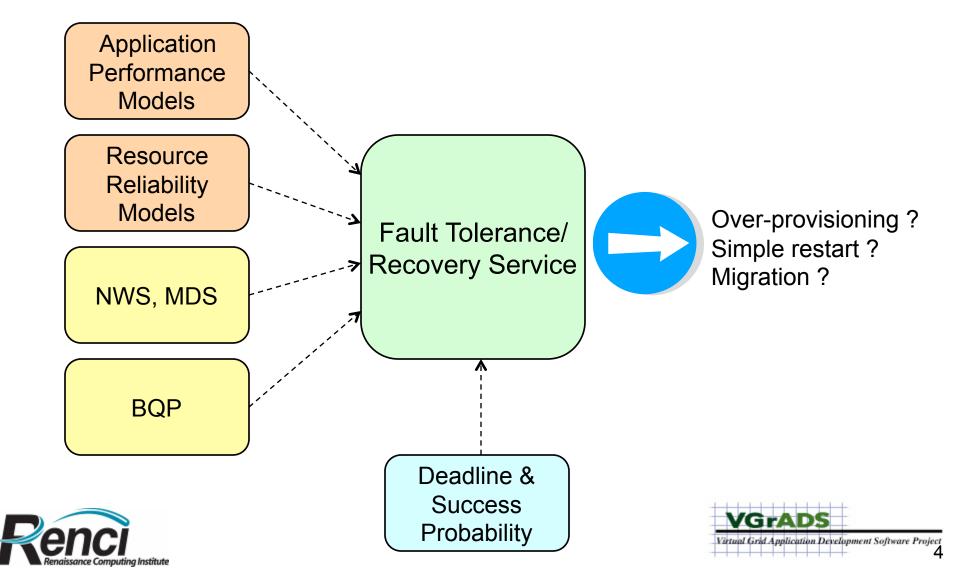
#### **Motivation**

- Reliability and performance are related
  - failure is the limiting case of poor performance
  - both involve measures of behavior over time
- Large, complex workflows are sensitive to failures
  - faults are the norm, rather than exceptions
    - distributed systems, services and resources
  - completion "guarantees" are problematic
    - workflow completion is probabilistic in the presence of faults
- Many time-critical workflows are deadline driven
  - severe weather events, disaster response, …

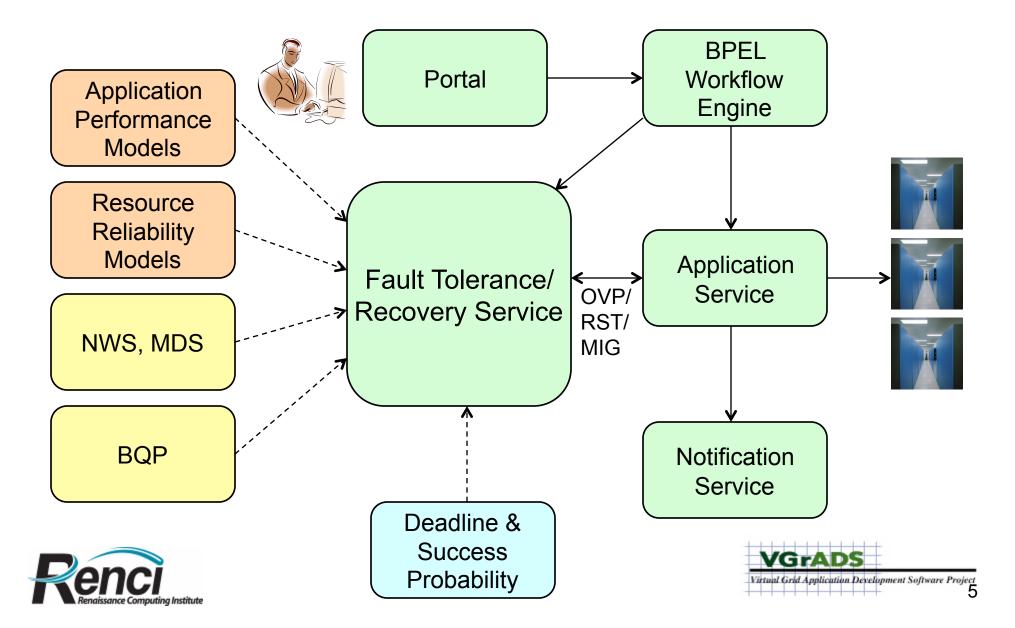




### Fault Tolerance/Recovery (FTR) Service



#### **Simplified Architecture**



# **FTR Algorithms**

#### **Notation** •

- p<sub>i</sub>: one hour failure probability of resource *i*
- *h<sub>i</sub>*: expected execution cost of application on resource *i* 
  - queue wait time
  - expected computation time
  - expected communication time
- x : required success probability
- d : required deadline
- If the reliability function is linear, probability of failure •  $m_i = \max(p_i * h_i, 1)$
- **Resource** *i* represents (queue, #nodes) combinations ullet





# **Over Provisioning**

- Find
  - degree and resources for over-provisioning
- Number of application copies
  - meet a deadline *d* with a success probability *x*
- Solve the following optimization problem

For given [1..M] resources, find a partition 
$$P = \{s_1, s_2 \dots s_n\}$$
 of [1..M] such that  
 $1 - m_{s1} * m_{s2} * \dots m_{sn} \ge x \land |P|$  is minimum  $\land min \{h_{s1} \dots h_{sn}\} \le d$   
  
Probability of failure Minimum number of resources meeting deadline







### **Simple Restart**

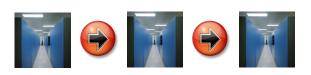
- Among the current available resources ...
  - (universe faulty resources)
- Find the best resource that
  - meets the deadline and success probability
- Mechanism can also be used for
  - runtime resource selection
- Two FTR mechanisms based on simple restart
  - restart with retries
  - restart if there is no progress







# **Migration**



- Find the best migration path
  - more complex optimization problem
  - need to optimize orderings
- Solve the following optimization problem

Find a partition  $P = \{s_1, s_2 \dots s_n\}$  and ordering  $\{s_1 > s_2 > \dots s_n\}$  such that

 $1 - m_{s1} * m_{s2} * \dots m_{sn} \ge x \wedge |P| \text{ is minimum } \wedge (t_{s1} + \dots t_{sn} + q) \le d,$ 

where  $t_{si} = (p_{si} h_{si} > 1)? 1/p_{si}$  and *q* is the migration overhead





### **Proposed Next Steps**

- Modeling resource reliability
  - draw data from job log files in the resources
  - maintain constant updates of reliability estimates
- Reliability Information Service (viz. BQP)
  - "What is the probability that a job, when submitted to a queue at a resource will die before completion because of a failure?"
- Balancing multiple mechanisms
  - given resource constraints
    - e.g., over-provisioning and migration
  - quantifying resource "wastage" with common metric





# **Proposed Next Steps (2)**

- Develop runtime rescheduling strategies
  - requires following changes in VGrADS software
    - ability to re-acquire slots
    - ability to relinquish unnecessary slots
    - ability to accommodate FTR directives
    - ability to schedule remaining workflow
  - requires following change in workflow engine
    - ability to run a re-scheduled workflow





# **Proposed Next Steps (3)**

- Deadline estimates
  - given a deadline for the entire workflow
    - find deadlines for individual applications
  - requires critical-path analysis





### **FTR Demonstration**

- Launch workflow from the LEAD portal
  - build new experiment
  - select forecast region and workflow
  - fetch data
- FTR service manages application execution by
  - choosing the fault-tolerance mode (OVP, RST)
  - invoking application service(s) as per current mode
  - monitoring for failures
- Workflow composer shows workflow progress
  - status of application currently running
  - workflow execution status via notifications etc.







#### S GETTING STARTED WITH THE LEAD PORTAL

Now that you are logged into the LEAD Portal, you have several additional capabilities available to you. Below is a list of links to these capabilities along with accompanying tutorials and documentation to help you get started using the LEAD Portal. More extensive documentation is available under the Help tab.

#### LEAD Portal Overview

• Take a tour of the LEAD Portal (requires Flash plugin)

#### My Workspace

- My Workspace provides a tree-based view of your projects, experiments, and data that are stored in MyLEAD.
- Take a tour of My Workspace (requires Flash plugin)
- Getting Started documentation

#### **Experiment Builder**

Experiment Builder provides the ability to create, manage, configure, and monitor LEAD experiments.

• Running a WRF Forecast Tutorial (requires Flash plugin)

#### Geographic Data Search

The Geographic Data Search tool allows you to perform searches for data by specifying any combination of spatial, temporal or data category parameters. You can access this tool without logging into the LEAD Portal, but now that you are logged in, you can also save data that you have found into your personal workspace.

• Take a tour of Geographic Data Search (requires Flash plugin)

COPYRIGHT © 2006 LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY, ALL RIGHTS RESERVED.

CONTACT US | LOGOUT













COPYRIGHT © 2006 LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY, ALL RIGHTS RESERVED.

POWERED BY





😇 LEAD Portal - Mozilla Firefox	
Ejle Edit Vjew Go Bookmarks Tools Help	
🖕 • 🛶 - 🥰 💿 🏠 🐲 https://portal-dev.leadproject.org/gridsphere/gridsphere/gridsphere?gs_mode=EDIT&cid=189	🔒 🗸 💿 😡 💽
🗋 Customize Links 🗋 Free Hotmail 🗋 Windows Marketplace 🗋 Windows Media 🗋 Windows	
ELEADPORTAL SPONSORED BY THE NATIONAL SCIENCE FOR LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY	INDATION
HOME MY WORKSPACE ABOUT LEAD DATA SEARCH EXPERIMENT VISUALIZE EDUCATION RESOURCES HEL	Ref spatialized in 22.19 kbw 2.4.12 kbw a serif and was well well well and well well well well well well well wel
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)	
Use the Fault Tolerant Recovery (FTR) service when submitting workflow	Submit
Back	

COPYRIGHT © 2006 LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY, ALL RIGHTS RESERVED.







Portal - Mozilla Firefox	
it View <u>Go</u> Bookmarks Iools <u>H</u> elp	
📫 - 🚰 区 🏠 🏠 https://portal-dev.leadproject.org/gridsphere/gridsphere?cid=1888gs_action=	<u>⊜</u> ♥ ∞ ⊂
omize Links 🗋 Free Hotmail 🗋 Windows Marketplace 🗋 Windows Media 🗋 Windows	
ELEADPORTAL SPONSORED BY THE NATIONAL SCIENCE FU	DUNDATION
E MY WORKSPACE ABOUT LEAD DATA SEARCH EXPERIMENT VISUALIZE EDUCATION RESOURCES HE	
luction Experiment Builder	-
Experiment Builder Portlet	
Experiment Wizard	
User: Test User 16 VG Project: Default Project	
Specify a name, description, and select workflow	
Name: test2	
test workflow	
Description:	
Workflow	
y Workflows (0) Sample Workflows (5)	E
VAM Initialized WRF V2.2 Forecast Description	
CrossCuttingConfigurations Config WRF_Static_V5_2_7_Preprocessor NAM_Initial_Conditions_V5_2_7_Interpolater	PS2WRF_V5_2_7_Interpolater WRF_Output_Files Config
ring data from portal-dev.leadproject.org	portal-dev.leadproject.org 🔒
	VGrADS Virtual Grid Application Developmen

Image: Solution of the state of t	AD Portal - Mozilla Firefox	
And a local of a local	Edit View Go Bookmarks Iools Help	
	Image: Second Se Second Second Sec	<u>2</u> © © G
The second of the lease because     The second of the lease because becaus	Customize Links 🗋 Free Hotmail 🗋 Windows Marketplace 🗋 Windows Media 🗋 Windows	
Operation         About LAD         Data SEARCH         Description         RESOURCE MED         RESOURCE MED           readulini         Experiment Builder         Experiment Builder         Experiment Builder         Image: Comparison of Search Sea		
bodubin Experiment Builder      Experiment Builder     Experiment Builder      Experiment Builder	LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY	
tandudini Epperiment Bulder		
	HOME MY WORKSPACE ABOUT LEAD DATA SEARCH EXPERIMENT VISUALIZE EDUCATION RESOURCES HELP	
Separate Wized         User: Text User 19 V0       Project: Default Project         Mann: text?       Description: text workflam:         Description: text workflam:       Description: text workflam:	troduction Experiment Builder	
Bit Part Part Part Part Part Part Part Par	Experiment Builder Portlet	80
here is it is dependence of the model coupler of th	Experiment Wizard	
Begind Skm         Regind Skm         Regind Skm         Begind Skm	Name: test2 Description: test workflow	
Region Type Selection         Or Regional Stam         Regional Stam         Regional Stam         Regional Stam         Or Regional Stam         Consus Staff Time         Dates and times in Greenwich Mean Time (GMT)         O Now (in other words, run a forecast using the most recent data available)         O Personal Stam         O Rate data available	Workflow: NAM Initialized WRF V2.2 Forecast	
	Model Domain Configuration	
Regional 20km         Orgenese Start Time         Determine in Greenwich Mean Time (GMT)         Own (in other words, run a forecast using the most recent data available)         Orgenese Start Time:         Orgenese Start Time:         Start Date:         Orgenese Start Date: <t< td=""><td>Region Type Selection</td><td></td></t<>	Region Type Selection	
Dates and times in Greenwich Mean Time (GMT) One win (in other words, run a forecast using the most recent data available) One set see specify: Surt Date:	🖸 Regional 20km	
<ul> <li></li></ul>	start Hour: 🕜 v	
<ul> <li></li></ul>	The satellite and satellite	Forecast Domain
Image: Sensitive Senstite Sensite Sensitive Sensitive Sensitive Sensitive S	Kelowinawna Lethoridge Mossea oRegina Winnipeg     Jawe Generation (Construction)     Jawe Generation     Jawe Generatio     Jawe Generation     Jawe Generatio     Jaw	center latitude: 38 0000
A Drag the balloon (?) to move the region. South Minnesolas Wisconsin Otione Oregon Idaho Wyoming Cregon Idaho Wyoming Cregon Idaho Wyoming California Vegas California	Victorian Val-d'Or	
Acting Coregon Idaho Wyoming Dakota Milwaukee Derot Bufaloe New York Vermont Settings Torology Derot Bufaloe New York Vermont New Hampshire Massachusetts Rhode Island Vert Vork Massachusetts Rhode Island Colorado Vert Virginia Vert Vork Vert Massachusetts Rhode Island Colorado Vert Virginia Vert Vork Vert Massachusetts Vert Massachusetts Rhode Island Colorado Vert Virginia Vert Virginia Vert Virginia Vert Vert Massachusetts Virginia Vert Vert Virginia Vert Vert Vert Vert Vert Vert Vert Vert	Washington         Montana         Dakota         Sudbury         Quebec         New           Portland         Portland         South Minnesota         Sudbury         Montral         Status	
Actificação esacramento veges Adazona	Oregon Idaho Wyoming Wigner Dakota Michigan Tenths Peterboreugh Vermont Sait	Settings
acifica Saramento Vitani Colorado Kanvas Missouri Louisviere Virginia an Esacramento Las California Vegas Alaquerque Angrilo Oklahoma Arransas Tennessee North Carolina Marviand Carolina Marves Virginia Connecticut I Radar Sites New Jersey Delaware New Jersey Delaware Marviand Marves Missouri Louisviere Marviand Marves Ma	En Derver Derver Uncon Kansan (Illinois Indiana Ohio / Philadelphile York Dadds Liberd	Map Layers
California Vegas     Vecnta     California Vegas     Vecnta     California Vegas     Vecnta     California Vegas     Vecnta     Vecnta     California Vegas     Vecnta	ncifica San esacramento Orani Colorado Kanyas Missouri Louisvillee Virginia Connecticut	
Losa 🗱 Arizona Now Altanta Altanta Altanta	California Vegas	
	Sebakersfields Audurenze Anne Oklanoma Arkansas Carolina Delaware	
		6
		VGrADS

#### **Questions**?



