LEAD-VGrADS meeting notes
November 3, 2005

Action items from the meeting:

1. Telecon for PIs – Dennis, Dan, Ken, Kelvin
2. Document resource requirements are for today's workflows (Dennis, Suresh)
3. What is the missing specification detail between LEAD workflow and vgDL? (Lavanya)
4. Use VGrADS calls for discussion on LEAD (Lavanya will co-ordinate)
5. Check on TIGRE (Chuck)
6. Work with TeraGrid (Dennis)
7. Revisit Virtual Grid API for LEAD application (Lavanya)
8. Ken will take a stab at an architecture diagram how VGrid can be accessed by LEAD type workflow.
9. Setup email list (Lavanya)

More discussions needed, other activities

1. Ability to use performance model tie, batch queue prediction from UCSB? (Alll)
2. Integrated system architecture
3. Fault Tolerance API for VGrid with workflows and SPMD
4. Scheduling in LEAD
5. MPI jobs on virtual grid., subdivide clusters, batch queues

Dan's opening points:
LEAD wants to take advantage of virtual grids. The most important aspect of LEAD is the BPEL based workflow orchestration. Multi-level Fault tolerance being worked at UNC e.g. checkpoint recovery based mechanisms complemented with at workflow level

Goal of meeting is to plan some co-ordinated research activities as well as target a LEAD-VGrADS demo for SC2006

Dennis:
Couple of holes in LEAD planning by design - dynamic resource brokering, scheduling, fault tolerance (dynamic behavior), resource scheduling. This is place for VGrADS. Need fault tolerance at multiple levels - service failures, unavailability of resources, integrate accurate histories of every step of execution, there is metadata that ties in with virtual data system. Identify failures and manage scheduling policies

Adaptation lessons from GrADS – Need to be careful or it can be too complicated

VGrADS Overview - Chuck:

Vision for the future is scientist doesn't have to make phone calls. Today it is bits and pieces. Can we build an API that virtual grid can be leveraged?

Lessons from GrADS: Mapping and Scheduling for MPI - Grid hard for MPI

Performance model construction is hard, automate the really hard things.

Partially automated the scheduling, fault tolerance is not built into virtual grids yet. Static parts of the virtual grids are what is there

GrADS had an extensive launch mechanism that has not been ported back completely Scheduling workflow task graphs on the virtual grid. We don't have capability of supporting readjustment if virtual grid disappears.

What does application need to do?

VGrADS speak about qualitative constructs to specify resources such as LooseBag, TightBag, Cluster.

Note: Definitions from the VGrADS document

> *LooseBagOf: Set of heterogeneous processors with possibly "poor" connectivity*
> *TightBagOf: Set of heterogeneous processors with "good" connectivity*
> *Clusterof: A set of homogeneous processors with "good" connectivity*

What is missing in vgDL? - Data management in virtual grid API

LEAD Overview by Beth

Meteorological workflows pre-LEAD technology was a simple static, serial workflow. Get data from sensors, analysis/assimilation, run the models, and present it to the user. Goal of LEAD would be to make the loop in 30 minutes. The LEAD workflows are data driven. Interesting results from the data mining drive new sub workflows. Need to adapt to weather and workflow as well as resources.

What is the bottleneck in the workflow? The computational models is the longest running and takes the most time.

What is an ensemble? Still being determined from the science perspective. But usually is a collection of model runs with different inputs/physical conditions that result in variation in the output. Each member of the ensemble itself is an MPI job. Similar to a parametric study.

The goal of LEAD is to have an adaptive and dynamic system in the closed loop forecast situation. Need on-demand resource provisioning for each ensemble. The actual size of the run can be adjusted based on how many resources there are available. Different resolution of the data (grid spacing) can be used based on time to forecast and resource constraints. Need ability to redirect sensors towards interesting events. For example, CASA NEXRAD adaptive doplar radars can be steered. Ideally the system may start with data from a large grid, find the interesting events, grab data from a smaller, finer grain

regions and repeat. Scheduling constrained by data streaming Grid spacing changes observation mechanism, generating grids is hard

What is holdup for recreating runs – can we recreate a situation like Katrina? Experiment available – data is available, track of storm prediction. Communication is unavailable. During Katrina, it is not the radars that went out but the networks. Ken K makes the observation that you could have a real-time deadline. Which is true, but problem is it is not clear what it is. Conditions are changing in real-time.

What is experiment in LEAD?
It is a feedback loop – deadline driven loop. Interesting scheduling – fixed time, cover as many runs as you can. Start with deadline. How many resources do I need? Different kind of scheduling problem. Adapt computation as well. Complex scheduling problem – limited amount of time, most precision, don't necessarily know workflow ahead of time, inverse scheduling.

What is hard in this workflow? data searching, configuration hacking. Done manually today. There are a lot of community data products - data streams, catalogs, etc

What is the level that LEAD will like to specify for workflow? Can describe higher level description? Semantic content → for workflows … automatically divine something.

Discussion on LEAD Portal: Researcher sit down, choose data sources, select and execute workflows, The portal gives graphical interfaces for selecting the data sources and for creating the workflows. The workflows are composed in a modified version of BPEL developed at Indiana. All LEAD transactions are captures as events in the myLEAD for providence, error tracking.

Funding discussion:

Perspectives from Ken

- probably do something by virtue of money we have
- This year's focus in VGrADS is LEAD
- Goal for SC: Run a LEAD experiment on the VGrADS

Some is in LEAD funding as well. Dynamic workflows is target for this year.

Possibility of scheduling research problems

- Some problem of scheduling can be solved
- Not sure about resources - Anirban is going to graduate this year, 1 FTE at Rice.

In loop scheduling, There is dynamic quality in GridSAT as well.

What we need is to track a storm dynamically. Capitalize post-Katrina hysteria to highlight problems and seek additional funding. Rice did crisis management study – disaster response and came up saying they need exactly what LEAD is doing.

Need a national level plan about pre-empting cycles for lower priority sciences for such problems.

Dennis: LEAD's core vision is dynamic allocation of resources, not sure if money is there. TeraGrid wants to be able to do that. No one group is funded to do this.

Resources we have are not enough , can we make a case?
Need to show capability
TeraGrid batch queues

Show we have strategy through collaboration this year in Katrina situation. We use historical data to show how it will work

Dan: Money will have to come from some other place other than NSF.

Not sure if FEMA or Homeland security has money.
Problem with Houston was mis-predirected traffic related to mis-predicted weather

Chuck: Get earmarked money in the highway bill. What is time scale? What are mechanisms for scheduling that support?

NOAA is a possibility, DoD is a possibility

We are probably not ready to do it today. Steps to get us there: VGrADS's year of LEAD: Where we can get to a place where we can get some money?

Get a small group together to talk about funding – Dennis, Dan, Ken, Kelvin

More detailed discussions:

Lavanya showed examples of vgDL and the API for creating a simple workflow. 3 step workflow, doesn't have streaming data or anything She also showed an early diagram on LEAD-VGrADS integration that shows a LEAD-VGrid Manager that manages interaction between workflow engine and the virtual grid.

Ken showed the diagrams from the Pegasus interaction and how the components interacted with each other.

SC06 demo discussion:
Define a SC05 demo today and try to add dynamic characteristics and virtual grid

Suresh:

Walked us through the portal demo for SC05. User grabs data from GeoGUI to configure a workflows. Input determine input for the workflow. Workflow structure is static today.

Input data is determined as we go. Each service creation is submitted to the PBS Queue. Selection of resources is manual. Always runs on particular resources.

BPEL doesn't do mapping, but it is done before hand now but it needs to be done before each task to accommodate dynamic workflows. First launches all services before workflow, and then registers it with a registry. Once registered the BPEL engine will find them from the registry. Future more control, change location of service. Everything is submitted through GRAM.

One other LEAD Scenarios is to run these workflows for classrooms full of students. Education workflow where time is limited.

Global planning is goal of VGrADS. Dynamic rescheduling – schedule everything statically, build another workflow and schedule that. Can chop workflows strategically

Performance analysis and prediction strategy (John Mellor-Crummy's and Gab's work)
- run EMAN through, predictions on run-time on different workflow patterns, volume of data and latency between nodes. Doesn't work on MPI parts, do prediction of nodes. Works mainly on typical scientific codes. Small instances of WRF can run on a single processor.

Can you fake a multi-processor behavior from uniprocessor?
Probably can but need some work. Need some predictor for communication performance

Can we use Adolfy's methodology on WRF?

Conversation about getting data from previous runs. History of experiments - is execution time available? Yes it is available. Time will vary based on input data.

Can we try to build a model with this data from the LEAD experiments that is being stored now?

What do we do for SC06?

Recap benefits for each group.

What is LEAD expecting?
→ cut down time and support dynamic adaptive nature of resources, on demand scheduling
→ performance modeling
→ dynamic adaptive workflows

{→ commercial quality deployable software}

What interesting research does LEAD bring to VGrADS?
→ Scheduling strategies (to deadline, static/dynamic, streams)

Ken, Chuck: Can we leverage LEAD as a TIGRE?

IU will contact a resource broker
- we want to do these steps, build us a virtual grid
- need ability to contact batch queues such as TeraGrid batch queues

Can you run it without a gram-job sumit and directly run it?

Does virtual grid have knowledge of batch queues? No at this point.

Prototype system stages:

1st system
- simple job
- possibly a MPI job

2nd system
- static workflow as it is today

3rd system: Next step (dynamic condition)
Dynamic Resource mapping, Application level

4th system - Can we do it on TeraGrid?

UCSD is looking at service manager issues like advanced reservations, batch queues etc.

Can we do incremental scheduling from workflows? Today through allocation of multiple virtual grids

Need to remember that it is asynchronous communication of workflow. When job completes it will notify back

Possible Resources we can target:

SC 06 demos
- 1 cluster with 64 (Rice)
- UNC
- TIGRE (32/128 cluster)
- TeraGrid

Will vgES support batch queues?

Not realistic for SC06 demos – excessive adaptation especially for failures

BPEL has notion of exceptions, GPEL doesn't have. It is meant to be asynchronous event driven, parallel system. Maps to VGMon – in network queries, provides real-time information, online performance

Discussion of XML/RPC vs SOAP.

Notion of providence in LEAD system - derive data for modeling. Scientists can track it