Parameter Estimation using Scientific workflows

David Abramson, Colin Enticott, Tom Peachey
Faculty of Information Technology,
Monash University,
Clayton, 3800, Victoria, Australia
david.abramson@infotech.monash.edu.au

I. INTRODUCTION

In recent years there has been a great deal of interest in "scientific workflows". These allow scientists to specify large computational experiments involving a range of different activities, such as data integration, modelling and analysis, and visualization, to name a few. Activities can be composed, often using a graphical programming environment, so that the output of one stage can be passed as input to the next, forming a pipeline of arbitrary complexity. Scientific workflows have been used to great effect in a number of different disciplines including Computational Chemistry, Ecology and Bioinformatics.

Importantly, many workflow engines double as programming environments for the Grid. Whilst there is no standard ways of doing this, a number of engines effectively expose the Grid middleware APIs. For example, Kepler [3][4] exposes a variety of middleware layers, from Globus through to ad-hoc interfaces like SSH. Other engines, like Triana and Taverna allow users to invoke services as Web Services, but provide no explicit support for Grid middleware. In spite of its significant power, Kepler, and many other current workflow systems, do not support dynamic parallel execution of the workflow and its components. This means that users must explicitly code a workflow to cause it to execute elements in parallel. This significantly complicates the workflow and obscures the underlying business logic.

In another paper, we have shown how to augment Kepler with a Tagged Data Flow Architecture Director (TDA) [5]. This new system, called Nimrod/K, extends Kepler by providing powerful mechanisms for exposing and managing parallelism in the workflows, and this provides an ideal platform for using workflows for parameter estimation. Unlike the current Nimrod tool chain, Nimrod/K makes it possible to run sweeps over workflows, and workflows that contain sweeps. This novel approach leverages Kepler's power in building complex workflows, and Nimrod's ability to execute sweeps over grid resources.

II. THE CURRENT NIMROD TOOL SET

Over the past 15 years and prior to the wide promotion and adoption of workflow engines, we have constructed a tool set called Nimrod that automates parameter estimation techniques using the Grid[1][2]. One of the Nimrod tools, Nimrod/G, supports complete parameter sweeps which have been used very effectively, with examples in environmental modelling, bioscience research, engineering and chemistry, to name a few [7]. A user provides details of the parameters and computational tasks to be performed, and the system generates and runs all combinations of the parameter values. Nimrod/G, however, can also serve as a Grid job management system for other software, including the other members of the Nimrod family.

Although the favoured method for exploring the parameter space of a computational model is a full parameter sweep, may be impractical where there are many parameters, especially if the model is computationally intensive. The Nimrod/E tool[6], on the other hand, automates the design of fractional factorial experiments. Here, the user specifies the factors and which interactions can be ignored. Then the *nimrodFracDes* tool produces an efficient design, and generates the parameter values for the resulting jobs in a form suitable for Nimrod/G. When all jobs are complete, another tool, *nimrodFracAn*, produces analyses of the results for each output value of interest. This tool produces various outputs that show the relative size of the various main effects and interaction effects. For example, one of these outputs is a Lenth Plot, as shown in Figure 1. The Lenth plot shows effects in order of absolute size and horizontal lines giving significance levels. The tool also produces estimates of the results for a full parameter sweep, which may then be used by visualization software.

III. NEW KEPLER ACTORS

A. The Nimrod/G actor

As discussed, the current Nimrod tools provide parameter exploration tools to complement Nimrod/G's Grid execution and meta-scheduling capabilities. One of these tools performs a parameter sweep that creates a Grid job for each of the parameter combinations in the parameter space. We have created a `ParameterSweep` actor that uses the same parameter syntax as the Nimrod/G implementation to generate parameter combinations in a workflow. This actor works with the current PtolemyII directors such as Synchronous Dataflow (SDF) and Process networks (PN). However, when used in combination with the TDA director in Nimrod/K, parallel execution of the different parameter combinations allows the workflow to execute efficiently on the Grid. The `ParameterSweep` actors can be chained

978-0-7695-3535-7/08 $25.00 © 2008 IEEE
DOI 10.1109/eScience.2008.97

together to create more elaborate parameter combinations and also create sub-sweeps.

B. *Nimrod/E Actors*

The design actor for Nimrod/E is built using the same principles as the `ParameterSweep` actor and is named the `FractionalFactorialDesigner`. For the analysis section of the workflow, the functionality was broken into four actors, as seen in figure 1: The `EffectEstimator` collates the results of the sweep and multiplies by the analysis matrix to produce estimates of the effects. The `FullSweepInterpolator` uses the effect estimates to predict the results for a full parameter sweep. The `DanielPlotter` orders the estimates, performs a Gaussian transformation and display the resulting Daniel plot. The `LenthPlotter`, orders the estimates according to their absolute values, computes confidence limits and graphs the resulting points and limits as a Lenth plot.

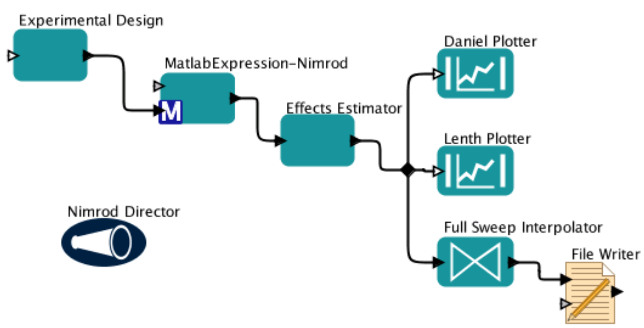

Figure 1. Experimental Design

IV. CASE STUDY

Mathematical models of the heart show considerable promise for understanding the underlying mechanisms and for clinical diagnosis and treatment. The model chosen concerns excitation and contraction in rabbit ventricular muscle cells [6]. Intracellular flows of calcium, sodium and magnesium ions were modelled as a system of ordinary differential equations using Matlab.

We first used the `ParameterSweep` actor to sweep over the nine factors, *A* to *J*. We chose two values for each factor, producing an experiment of 512 jobs. The second experiment used the experimental design functionality to further investigate the response function as shown in figure 1. The `FractionalFactorialDesign` actor was used to produce a design that would estimate the main effects and two-way interactions of these factors which created 128 jobs. Effects were estimated and used to produce Lenth (Figure 2) and Daniel plots and to predict results of a full sweep. All these experiments identified the interactions between the parameters.

We performed these two experiments on a shared local compute resource. During the parameter sweep, Nimrod/G obtained 105 processors and the Nimrod/E experiment obtained 58 processors before the experiment finish. This experiment has shown that Nimrod/K provides a natural and easy mechanism for specifying the use of parameter estimation techniques over existing workflows and is capable leveraging Grid middleware. All of the machinery developed to date is capable of sweeping across parameter combinations regardless of the complexity of the computational steps of the workflow. Thus, much more complex pre-existing workflows can be modified using our new actors to perform large-scale complex parameter estimation experiments.

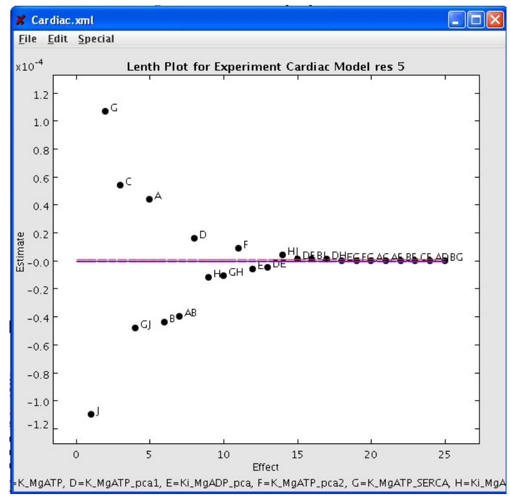

Figure 2. Lenth plot

ACKNOWLEDGEMENTS

The Nimrod/K project is supported by Australian Research Council ARC Discovery grant. We are grateful to our colleague, Ilkay Altinas, from the San Diego Supercomputing Center for her assistance.

REFERENCES

[1] D. Abramson, J. Giddy and L. Kotler, “High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid,” In *Int’l. Parallel and Distributed Processing Symposium (IPDPS),* Cancun, Mexico, May 2000. http://www.csse.monash.edu.au/~davida/nimrod/.

[2] http://www.messagelab.monash.edu.au/Nimrod

[3] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher and S. Mock, “Kepler: Towards a Grid-Enabled System for Scientific Workflows,” in the *Workflow in Grid Systems Workshop in GGF10* - The 10th Global Grid Forum, Berlin, March 2004.

[4] http://kepler-project.org

[5] Abramson, D., Enticott, C and Altinas, I. “Nimrod/K: Towards Massively Parallel Dynamic Grid Workflows”, to appear IEEE Supercomputing 2008, Austin, Texas, November 2008.

[6] S. Amirriazi, S. Chang, T. Peachey, D. Abramson and A. Michailova, *Optimizing Cardiac Excitation-Metabolic Model By Using Parallel Grid Computing,* Biophysics 2008, Long Beach, California, February 2008.

[7] http://www.messagelab.monash.edu.au/EScienceApplication