Trident: Scientific Workflow Workbench for Oceanography

R.S. Barga[1], J. Jackson[1], N. Araujo[1], D. Guo[1], N. Gautam[1], K. Grochow[2], E. Lazowska[2]
Microsoft Research, Microsoft Corp.[1] Dept. of Computer Science, University of Washington[2]
barga@microsoft.com

Abstract

We introduce Trident, a scientific workflow workbench that is built on top of a commercial workflow system to leverage existing functionality. Trident is being developed in collaboration with the scientific community for oceanography, but the workbench itself can be used for any science project for scientific workflow.

1. Introduction

We are at the dawn of a revolutionary new era of "e-science." Advances in technology are transforming discovery in all scientific fields, in two important ways. First, massive experiments are being carried out by simulating the real world using computers with thousands of processors. Second, large numbers of sensors are being deployed to gather data on the sea floor [1], on glaciers in the Swiss Alps [2], and across the landscape on which we walk [3]. These approaches share a common trait: they produce enormous amounts of data that must be captured, transported, stored, accessed, visualized and interpreted to extract knowledge. This *computational knowledge extraction* is at the heart of 21st century discovery.

Scientific workflows are proving to be the preferred vehicle for computational knowledge extraction and for enabling science at a large scale. Workflows provide a scientist with a useful and flexible method to author complex *data analyses pipelines* composed of heterogeneous steps ranging from data capture from sensors or computer simulations, to data cleaning, to transport and storage, and provide a foundation upon which results can be analyzed and validated. However, building and maintaining robust scientific workflows systems is proving to be extremely costly, and the long term sustainability of academic research prototypes is an open question. In our work we evaluate how a scientific workflow workbench can be implemented on top of a commercial workflow enactment engine, specifically Windows Workflow, to leverage existing functionality. Trident only implements functionality and services required for scientific workflow management. In doing so it offers a robust platform for science groups to spend more time on their science and less time writing code.

2. Trident for Project NEPTUNE

NEPTUNE is the first Regional Cabled Observatory, on the Juan de Fuca plate off the coast of Washington. NEPTUNE will place thousands of chemical, geological and biological sensors on 2000 kilometers of fiber optic cable on the sea floor, streaming data back to shore for analysis. NEPTUNE will transform oceanography from a data-poor to a data-rich science. It will help unlock secrets about the ocean's ability to absorb greenhouse gases, and about how stresses on the seafloor cause earthquakes and tsunamis along Pacific coastlines.

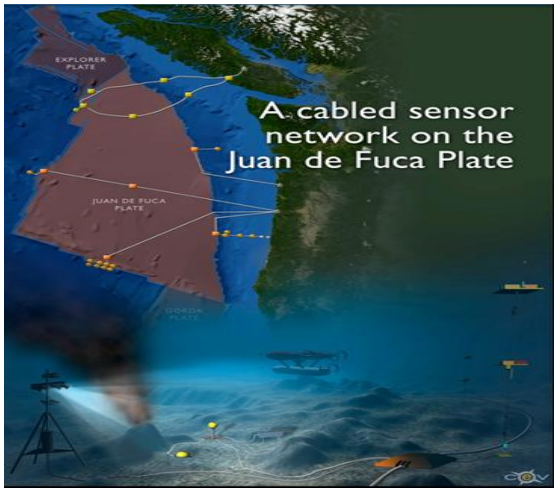

Figure 1 – Regional Cabled Observatory

Trident is being developed as part of a collaborative project between The University of Washington and the Microsoft Technical Computing Group [5], to provide Project NEPTUNE with a scientific workflow workbench. Trident, implemented on top of Windows Workflow Foundation, allows scientists to explore and visualize oceanographic data in real-time and provides an environment to compose, run and catalog workflows. Trident uses Silverlight [6], a freely downloadable cross-browser, cross-platform, and cross-device plug-in for delivering .NET based applications over the Web, so a scientist using Windows, MacOS or Linux can use Trident to compose and run experiments from any web browser.

Other features in Trident for data intensive research include: automatic provenance capture, cost estimates of the resources a specific workflow will require, automatic scheduling of workflows on HPC clusters, and support for fault-tolerance and recovery from failures.

3. What is demonstrated?

We demonstrate how Trident addresses various real-world issues that oceanographers encounter in attempting to turn a sea of data that is streaming from sensors in the ocean into visualizations and data products to support their research. In addition, we outline exactly how we leverage a commercial workflow enactment engine to support scientific workflows. Highlights of the demonstration are described in the remainder of this submission, but this list of features we will demonstrate is not exhaustive but limited by the page limit.

3.1 **Easy and rapid ad-hoc workflow design**

We will demonstrate visual programming of workflows using a catalog of existing activities and complete workflows, using only a web browser. The authoring of a workflow is an important aspect of any workflow system, as it allows the researcher to specify both steps and control dependencies in the data analysis pipeline. Easily finding and adapting an existing workflow is key to effective workflow prototyping. As the end users for Trident are not seasoned programmers, it offers a graphical interface that enables visual programming, as well as a web based portal for authoring and launching workflows via a browser. Trident also provides a tiered library that hides the complexity of different workflow activities and services for ease of use.

3.2 System wide registry for sensors to services

We also demonstrate the Trident registry, which can include all objects of interest for Project Neptune, ranging from sensors in the ocean, services providing access to data, to workflows, and *versioned* results. There are an increasing number of tools and databases in the sciences available as a Web Service. As a result, researchers are not merely faced with a data deluge but also face a data *source* and *service* deluge and need a tool to organize, curate and search for services of value to their research. The Trident registry enables researchers to search on tags, keywords and annotations to see what services are available. Semantic tagging enables researchers to find a service based on what it does, or meant to do, and what it consumes as inputs and produces as outputs. Annotations allow the researcher to understand how to operate and configure it, and the registry records the version history.

3.3 Visualization of oceanographic data

One of the primary goals of Trident is to convert raw sensor data into useful data products, in particular visualizations. COVE [7] is a tool that provides visualization of ocean data. We will demonstrate how an oceanographer can use COVE to create on demand visualizations by invoking workflows on Trident. Together, Trident and Cove make up what Jim Gray referred to as "The Ocean Scientists' Workbench" to enable collaborative ocean science.

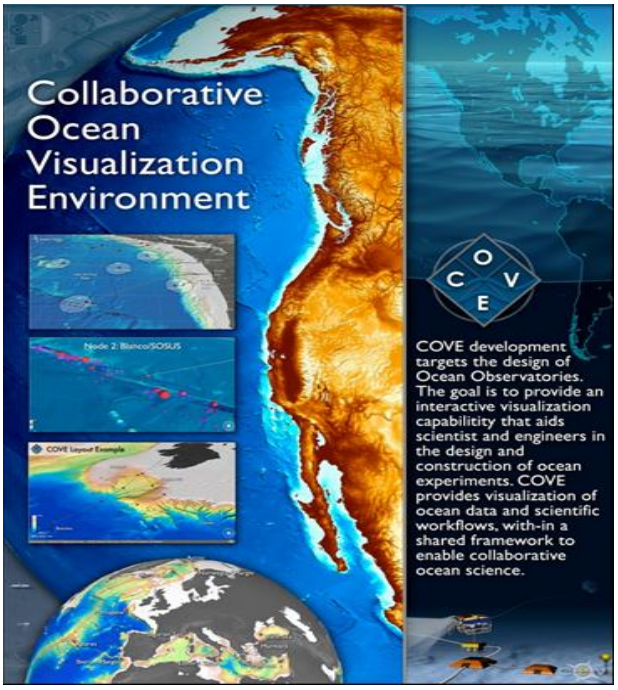

Figure 2 – COVE Interactive Visualization of Ocean Data.

4. REFERENCES

[1] Project Neptune http://www.neptune.washington.edu/.

[2] Swiss Experiment, http://www.swiss-experiment.ch.

[3] Life Under Your Feet http://lifeunderyourfeet.org.

[4] Microsoft Windows Workflow Foundation (WinWF) http://en.wikipedia.org/wiki/Windows_Workflow_Foundation.

[5] Technical Computing Group of Microsoft Research http://www.microsoft.com/science.

[6] Microsoft Silverlight http://silverlight.net/.

[7] COVE Oceanographic Visualization Workbench http://www.cs.washington.edu/homes/keithg/oceans.html.