SI2-SSE: Pipeline Framework for Ensemble Runs on the Cloud
Beth Plale (PI), Indiana University | Craig Mattocks (Co-PI), University of Miami

Get Involved:
• Project Website: [http://d2l.indiana.edu/cloud-for-climate](http://d2l.indiana.edu/cloud-for-climate) with links to Sigiri/software releases, publications, news, and updates.
• Join our monthly call to become involved with the project and efforts to develop/assemble the tooling. First Wednesday of each month, January – May 2013, 1:00 PM EST (New York Time) 1-800-940-6112 Pin: 000922# (see website for future call times/dates)

Goals: The goal of the project is to develop and/or assemble tooling to enable rapid and cost effective deployment of anywhere between 500 and 15,000 small jobs on cloud resources. This project develops a pipeline framework for running ensemble simulations on the cloud; the framework has two key components:

Ensemble deployment: we develop/test tooling for rapid and cost effective deployment of anywhere between 500 and 15,000 small jobs on cloud resources. The specific motivating application is a storm surge model called SLOSH (Sea, Lake and Overland Surges from Hurricanes). The current cloud platform is Windows Azure cloud.

Metadata harvest: we can reduce the burden of sharing scientific datasets resulting from the use of cloud resources through automatic metadata and provenance capture. Ensemble simulations result in complex data sets, whose reuse could be increased by expressive, granule and collection level metadata, including the lineage of the resulting products, to contribute towards trust.

Through this project users can analyze a storm surge with a larger number of SLOSH instances to measure the affected region with a higher precision. The primary metrics are total execution time (maximum delay) and cost.

Status/Current Approach: We assume all jobs must complete and that a user is able to effect a specific trade-off between cost and delay metrics. The load skewness and I/O overheads are two metrics affecting both the cost and the execution delay.

Current approach:
→ Execute large number of SLOSH instances in cloud by leasing resources based on user demands
→ Submit SLOSH jobs using web service
→ Service manager within cloud parallelizes jobs by applying load balancing, reducing storage and I/O overheads in handling intermediate results, and aggregating intermediate outputs. Use two approaches: a MapReduce runtime (Twister4Azure) and a Sigiri Middleware.

Current efforts include:
→ Feasibility analysis of a number of middleware candidates for supporting the computational loads. (complete see [1] below)
→ Batch Execution of SLOSH ensembles with balancing loads and elastic processing of jobs through revising the leased resources in an on-line fashion. (on going)

Anticipated Impact: Users who could benefit from the application include the National Hurricane Center who is a partner on the project, Federal Emergency Management Administration (FEMA), the U.S. Army Corps of Engineers, and state and local emergency managers. The components developed as part of the pipeline are applicable across a number of scientific applications include climate monitoring and simulation, processing data from satellites, processing data from a very-long-baseline interferometry.

Broader impacts: Tools of this project are incorporated into a graduate course, Scientific Data Management and Preservation, where students from other disciplines run their applications in the pipeline framework, into workshops we host on the framework, and through active seeking of applications.