Elastic Computation for Variable-Demand Streaming Workflows
Position Paper for NSFCloud Workshop on Experimental Support for Cloud Computing

Paul Heinrich, Merriam-Powell Center for Environmental Research, Northern Arizona University
JD Knapp, Real-Time Intelligent Systems and Networks Laboratory, NAU
Souparno Ghosh, Dept. of Mathematics and Statistics, Texas Tech University
Paul Flikkema, Informatics & Computing Program, NAU (paul.flikkema@nau.edu)

Today, the bulk of cloud-based computational loads can be characterized as batch processing. However, streaming data applications are becoming a rapidly-growing fraction of the global networking and computational load. While this increase has so far been largely driven by financial analytics and personal device applications, demand for complex analytic processing will also arise from cyber-physical systems, including wide-scale internet-of-things services for research and as well as business and retail markets. Many of these applications will require data-driven inference of complex dynamical system models, which in turn will be characterized by rapid, event-driven transitions to new modes with different models. For example, agents controlling power flows in the smart grid may be required to learn, and adapt to, jump changes in network topologies due to failures or step functions as new resources come online. In environmental sensing (e.g., for prediction of flooding), soil moisture dynamics are characterized by complex nonlinearities (e.g., water saturation) [1] and are punctuated by pulses of rain events, requiring rapid model switching and adaptation. In these applications, computationally-intensive machine learning techniques can be used to infer, evaluate, and select, and calibrate competing models.

We are interested in exploring computational frameworks for variable-demand workflows. The cloud paradigm could enable highly efficient use of distributed computational resources, combining local, dedicated data centers handling quiescent or steady-state loads with the ability to quickly marshal and release cores and storage in the cloud, triggered by sensed events that signal modal changes in complex systems.

Background: We have developed and are currently deploying distributed cyberinfrastructure for the Southwest Experimental Garden Array (SEGA), an NSF-funded facility that will support experimental research on the effects of climate change on plant communities and ecosystems [2]. The SEGA CI interconnects wireless sensor/actuator networks in twelve geographically-distributed gardens and a 48-core Real-Time Data Center using DataTurbine streaming data middleware [3], supporting management of sensing, networking, inference and control of plant water availability. We have recently initiated the development of a DataTurbine-based architecture for streaming data workflows based on *processors*, software abstractions of (i) blocks in feedback control systems or (ii) nodes in generalized graphs of distributed processing systems [4].

Objectives: We propose to use CloudLab and/or Chameleon in experiments to perform on-line data-driven inference of dynamic soil moisture models. The challenges we will

explore are (1) how to design cloud-based CI for rapid response to variable workloads (this is currently an area of active research in the community comparing containers and virtual machines); (2) how to integrate cloud resources into flexible streaming data processing. In essence, our long-term goal is to develop approaches that allow the rapid scaling up/down of parallel processing power to meet event-driven, short-term, and time-sensitive demands for machine learning.

Approach: We will leverage our experience using DataTurbine for creation, management, and processing of streaming data to develop preliminary experiments that explore responsiveness and management issues in both containerized and VM environments. We will study how to extend our Processor Control Library (PCL) [4] to support rapid development of on-line algorithms for model inference and state estimation that can marshal and release cores based on inference needs, latency requirements and available cloud resources. These experiments will entail development of simulated streaming sources (based on models of soil moisture dynamics) and adaptation of existing algorithms for model inference [1], as well as working with CloudLab and Chameleon researchers in the design of layered approaches for requesting, using, and releasing cloud resources at high levels of responsiveness while maintaining efficient use of those resources.

References Cited

1. S. Ghosh, D.M. Bell, J.S. Clark, A.E. Gelfand, and P. Flikkema, "Process Modeling for Soil Moisture Using Sensor Network Data", Statistical Methodology (Special Issue on Modern Statistical Methods in Ecology), vol. 17, pp. 99-112, Mar. 2014.

2. P.G. Flikkema, K.R. Yamamoto, S. Boegli, C. Porter and P. Heinrich, "Towards Cyber-Eco Systems: Networked Sensing, Inference and Control for Distributed Ecological Experiments", IEEE International Conference on Cyber, Physical and Social Computing, Nov. 2012.

3. J. Shaeffer, J.D. Knapp, M. Miller, and P.G. Flikkema, "A Middleware-Based Approach to the Design of Interconnected Sensor/Actuator Networks", 2014 IEEE International Workshop on Real-Time Cyber-Physical Systems, June 2014.

4. J. Knapp, M. Elo, J. Shaeffer, and P.G. Flikkema, "Towards Intelligent Closed-Loop Workflows for Ecological Research," Dynamic Data-Driven Environmental Systems Science Conference (DyDESS), Nov. 5-7, 2014, Cambridge, MA (accepted).