POS: Online Learning for Memory-Aware Scheduling of Scientific Workflows

Carl Witt
Humboldt-Universität zu Berlin
Berlin, Germany
wittcarl@informatik.hu-berlin.de

Dennis Wagner
Humboldt-Universität zu Berlin
Berlin, Germany
wagnerde@informatik.hu-berlin.de

Ulf Leser
Humboldt-Universität zu Berlin
Berlin, Germany
leser@informatik.hu-berlin.de

A scientific workflow is a set of programs with DAG-structured data dependencies used to analyze large amounts of data. The execution of workflows often requires scheduling of thousands of tasks with heterogeneous resource requirements on distributed compute resources. To effectively utilize compute resources, schedulers require task resource usage estimates. We propose a predictive online scheduling (POS) approach that learns those estimates during workflow execution using online models.

Traditionally, research either focuses on predicting task resource usage [1] or on scheduling under the assumption of perfectly accurate task resource usage estimates [2]. Given the difficulty of the prediction task, as reported in the former branch of research, assuming complete information about task runtimes and communication costs for scheduling is not realistic. Task resource usage is subject to many factors, such as resource heterogeneity, input size, and background load. We propose an integrated approach that combines scheduling and resource usage estimation by adding learning about task resource usage to the scheduling objectives. This leads to a completely new scheduling scenario: scheduling under *partially controllable quality of resource usage estimates*. Compared to the traditional scenarios of perfectly accurate estimates or bounded estimation errors, this leads to a more complex but also more realistic execution model.

A second problem is that the majority of scheduling algorithms assumes infinite resources for resource types other than CPU. Memory is seldomly considered [3], but often a bottleneck resource when running complex analyses on large data sets [4], [5]. In addition, big data analysis frameworks like Hadoop require users to specify the required amount of CPUs and RAM per task, requiring more complex schedulers.

Our predictive online scheduling (POS) approach addresses both problems. It exploits that programs often exhibit learnable resource consumption behavior given their input size. Based on this observation, POS learns task characteristics in a pure online manner during workflow execution (Figure 1). Our approach can learn arbitrary characteristics of tasks; in our case study, we demonstrated its effectiveness by predicting peak memory usage of tasks. POS is based on three principles: 1) collect resource usage measurements of new tasks early, 2) prioritize tasks that have failed before, and 3) favor well-predictable tasks. The key idea is to collect new information

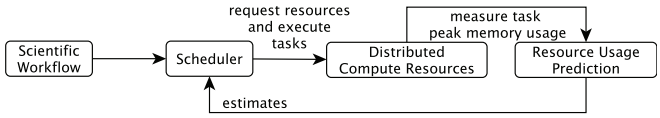

Fig. 1: Integrated scheduling and prediction.

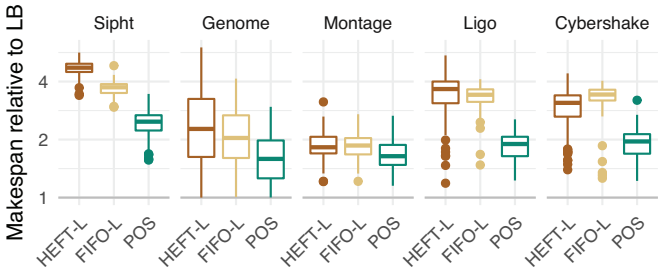

Fig. 2: Comparison of workflow execution times relative to a lower bound for different schedulers.

carefully, limiting the degree of parallelism of tasks with uncertain resource usage as long as resources can be utilized with tasks that have more predictable resource usage. This reduces memory wastage due to overestimating memory usage and underestimation, which leads to task failure.

We evaluated POS on a test suite of 1000 workflows (Montage, Sipht, etc.) generated by the Pegasus workflow generator [6]. We added randomly sampled memory usage for the tasks in the workflows and measured the impact of different scheduling policies and prediction models (linear regression, kernel regression, prediction models with controllable learning rate) on workflow execution time and memory utilization. Simulation experiments were conducted in DynamicCloudSim [7]. The accuracy of the simulation was validated by running a selection of workflows on real Hadoop clusters.

We showed that our scheduler mitigates memory bottlenecks which leads to significantly faster workflow execution. Figure 2 shows the distribution of makespans achieved by POS and learning versions of the FIFO and HEFT [8] schedulers that use the same prediction method as POS to predict future task memory usage. The more careful scheduling of POS leads to fewer prediction errors and thus faster workflow execution.

978-1-5386-9156-4/18/$31.00 ©2018 IEEE
DOI 10.1109/eScience.2018.00120

REFERENCES

[1] C. Witt, M. Bux, W. Gusew, and U. Leser, "Predictive Performance Modeling for Distributed Computing using Black-Box Monitoring and Machine Learning," *arXiv.org*, May 2018.

[2] M. A. Rodriguez and R. Buyya, "A taxonomy and survey on scheduling algorithms for scientific workflows in IaaS cloud computing environments." *Concurrency and Computation: Practice and Experience*, vol. 29, no. 8, 2017.

[3] R. Grandl, S. Kandula, S. Rao, A. Akella, and J. Kulkarni, "GRAPHENE - Packing and Dependency-Aware Scheduling for Data-Parallel Clusters." *OSDI*, 2016.

[4] A. Rheinländer, M. Lehmann, A. Kunkel, J. Meier, and U. Leser, "Potential and Pitfalls of Domain-Specific Information Extraction at Web Scale," in *SIGMOD Conference*. Humboldt-Universitat zu Berlin, Berlin, Germany, 2016, pp. 759–771.

[5] M. Bux, J. Brandt, C. Witt, J. Dowling, and U. Leser, "Hi-WAY: Execution of Scientific Workflows on Hadoop YARN," in *International Conference on Extending Database Technology*, 2017.

[6] G. Juve, A. L. Chervenak, E. Deelman, S. Bharathi, G. Mehta, and K. Vahi, "Characterizing and profiling scientific workflows." *Future Generation Computer Systems*, vol. 29, no. 3, pp. 682–692, 2013.

[7] M. Bux and U. Leser, "DynamicCloudSim: Simulating heterogeneity in computational clouds," *Future Generation Computer Systems*, 2015.

[8] H. Topcuoglu, S. Hariri, and M.-Y. Wu, "Performance-effective and low-complexity task scheduling for heterogeneous computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 3, pp. 260–274, Mar. 2002.