Contribution ID: **43** Type: **not specified**

Non-Stationary Gradient Descent for Optimal Auto-Scaling in Serverless Platforms

Friday, June 21, 2024 11:00 AM (30 minutes)

To efficiently manage serverless computing platforms, a key aspect is the auto-scaling of services, i.e., the set of computational resources allocated to a service adapts over time as a function of the traffic demand. The objective is to find a compromise between user-perceived performance and energy consumption. In this paper, we consider the "scale-per-request" auto-scaling pattern and investigate how many function instances (or servers) should be spawned each time an unfortunate job arrives, i.e., a job that finds all servers busy upon its arrival. We address this problem by following a stochastic optimization approach: taking advantage of the ability to observe the system \emph{state} over time, we develop a stochastic gradient descent scheme of the Kiefer–Wolfowitz type. At each iteration, the proposed scheme computes an estimate of the number of servers to spawn each time an unfortunate job arrives to minimize some cost function. Under natural assumptions, we show that the sequence of estimates produced by our scheme is asymptotically optimal almost surely. In addition, we prove that its convergence rate is $O(n^{-2/3})$ where n is the number of iterations.

From a mathematical point of view, the stochastic optimization framework induced by auto-scaling exhibits non-standard aspects that we approach from a general point of view. We consider the setting where a controller can only get samples of the transient – rather than stationary – behavior of the underlying stochastic system. To handle this difficulty, we develop arguments that exploit properties of the mixing time of the underlying Markov chain. By means of numerical simulations, we validate the proposed approach and quantify its gain with respect to common existing scale-up rules.

Primary authors: JONATHA, Anselmi (Inria); Dr GAUJAL, Bruno (Inria); Dr REBUFFI, Louis-Sebastien (Univ. Grenoble Alpes)

Presenter: Dr GAUJAL, Bruno (Inria)

Session Classification: Parallel session: Reinforcement learning for real-life applications