Proposition de stage Master 2ème année:

Optimal Design of Autoscaling Policies for Serverless Applications

Supervisor: Jonatha Anselmi (INRIA)
http://polaris.imag.fr/jonatha.anselmi
Grenoble, Spring 2024

Keywords: Applied probability, optimization, Markov decision process, Autoscaling, Cloud systems
Lab: Laboratoire d'Informatique de Grenoble (LIG), http://www.liglab.fr
Team: POLARIS, https://team.inria.fr/polaris (head: Arnaud Legrand)
Duration: 5 or 6 months
Language: English or french

Technologic Context

Load balancing is the process of distributing work units (jobs) across a set of distributed computational resources (servers). Exogenous jobs join the system over time through one or more dispatchers, and these route each of them to one out of N parallel servers for processing immediately upon their arrival. Then, each job leaves the system upon service completion at its designated server. Given the stringent latency requirements of modern applications, breaches of which can severely impact revenue, load balancing techniques are usually designed to optimize delay performance, and popular examples are Power-of-2 and Join-the-Idle-Queue (JIQ) [1]. Closely related to load balancing, *auto-scaling* is a term often used in cloud computing to refer to the process of adjusting the available service capacity automatically in response to the current load [4]. In this context, auto-scaling techniques are meant to control capacity (N) over time to avoid performance degradation, which yields unacceptably large delays, and overprovisioning of resources, which yields high infrastructure and energy costs. Google Cloud Run, Amazon Elastic Compute Cloud (EC2), Microsoft Windows Azure and Oracle Cloud Platform are examples of platforms that offer auto-scaling and load balancing features. Users of these platforms deploy their applications defining how the system should scale up resources in front of an increased load. Modern auto-scaling mechanisms are extremely reactive in the sense that they adapt capacity relying on fresh observations of the system state rather than historical data. This especially holds true in *serverless computing platforms*, or Function-as-a-Service, which nowadays provide the convenient solution to deploy any type of application or backend service [2].

Scientific Challenge

The load-balancing and autoscaling processes influence each other and can be coupled together to form a single controller. This is in fact the case in serverless platforms such as Google Cloud Run and Amazon Elastic Compute Cloud (EC2). The main goal of this internship is to design highly-scalable autoscaling control rules that are able to optimize a trade-off between performance (average delay) and energy consumption. This is tackled by formulating the problem in the framework of Markov decision processes (see, e.g., [3]), and the scientific challenge consists in deriving structural (mathematical) properties.

Requirements

The intern will have a solid background in applied probability and optimization, and a first course level knowledge about Markov decision processes in discrete time. The intern will also have some ability to write computer programs (in any language).

Location and Contact

The intern will be hosted in the POLARIS team. The POLARIS team is a joint team between Inria and LIG (Grenoble Computer Science Laboratory) and is located in Grenoble University main campus (https://batiment.imag.fr). For more information, please contact jonatha.anselmi@inria.fr.

References

[1] Knative Load balancing. https://knative.dev/docs/serving/load-balancing/, 2022. Online; accessed: 2022-04-04.

[2] N. Mahmoudi and H. Khazaei. Performance modeling of serverless computing platforms. *IEEE Transactions on Cloud Computing*, pages 1–1, 2020.

[3] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[4] C. Qu, R. N. Calheiros, and R. Buyya. Auto-scaling web applications in clouds: A taxonomy and survey. *ACM Comput. Surv.*, 51(4), July 2018.