Laboratoire de l'Informatique du Parallélisme ROMA team
(UMR CNRS - ENS Lyon - UCB Lyon 1 - INRIA 5668)

internship proposal - Master 1/2

Scheduling Replica Requests in Key-Value Stores

Loris Marchal, Sonia Ben Moktar and Étienne Rivière.

Keywords: Distributed systems, scheduling, algorithms.

Duration: 4 to 6 months, between January and July 2019.

Advisors:

Loris Marchal: CNRS researcher, LIP laboratory at ENS Lyon, France
(http://perso.ens-lyon.fr/loris.marchal/, loris.marchal@ens-lyon.fr)

Sonia Ben Mokhtar: senior CNRS researcher, LIRIS laboratory at Univ. Lyon 1, France
(https://sites.google.com/site/soniabm/, sonia.benmokhtar@insa-lyon.fr)

Étienne Rivière: professor at UC Louvain, Belgium
(https://uclouvain.be/repertoires/etienne.riviere, etienne.riviere@uclouvain.be)

Research environment

This internship will take place in the **ROMA team of the "Laboratoire de l'Informatique du Parallélisme" at ENS Lyon**, under the supervision of Loris Marchal, whose research interests include scheduling design for High Performance Computing platforms. The internship will also be co-advised by Sonia Ben Mokhtar (at LIRIS laboratory, Univ. Lyon 1) and Étienne Rivière (at UC Louvain, Belgium), both specialists of distributed systems and in particular distributed storage systems. There is a good **opportunity for the intern to spend one month in UC Louvain** at the end of the internship, provided the intern is interested (specific funding may be available for this).

Context of the internship

Many distributed systems have been proposed to optimize information storage and retrieval, in particular key-values stores such as Apache Cassandra or MongoDB. Even on well-provisioned systems, avoiding latency, that is, minimizing the delay of all data requests, is still challenging because of the unbalanced load among servers. In particular, the 5% slowest requests may experience much larger latencies than others, which is known as the *"tail latency"* problem. Data replication is commonly used to overcome this problem, which, in turn, asks for a good replication selection algorithms. When data has heterogeneous sizes, requests scheduled behind other requests for large data (and thus with large service time) may also suffer an additional delay: this is the *"head-of-line blocking"* problem.

Héron [1] is a replica selection algorithm recently proposed to overcome this problem. It recognizes requests for large values using bloom filters to avoid scheduling other requests behind them. Experiments show that it outperforms state-of-the-art algorithms and reduces both median and tail latencies.

However, we do not know if such a replication selection algorithm could be further optimized. More precisely, we would like to know how far is Héron from an optimal selection algorithm. A natural way to answer this question is to abstract the problem and to model the distributed system by selecting keys characteristics (while simplifying others) in order to get a simpler but tractable scheduling problem. In fact, optimal or guaranteed algorithms have been designed for similar problems in the scheduling literature: for homogeneous request sizes with total replication, it is known that the FIFO policy gives optimal results both for average and maximum response time minimization [3]. Data locality given by the replication policy may be precisely modeled using "restricted availabilities" as proposed in [2].

Objective of the internship

The objective of the internship is both to propose tight lower bounds on the achievable tail latency for the replica selection problem based on scheduling studies, and to identify which characteristics of the workload are crucial to design efficient replication strategies.

We have already identified two noteworthy specificities of the problem whose use in scheduling studies to design better bounds and algorithms would be pioneering:

- Thanks to the dynamic monitoring and scaling of the platform, the average load of each server is kept between 50% and 80%. This excludes extreme cases such as under-loaded or overloaded scenarios.
- The popularity of data items can be learned during the execution. Popular items receive much larger volume of requests, which may be scheduled specifically.

The internship consists into two parts, whose respective weight can be adapted to the student's skills and inclination:

1. Theoretical studies to model the problem, explore related scheduling works, design scheduling and replication selection algorithm, prove algorithm optimality or approximation ratios;
2. Experimental studies based on the statistical characterisation of workloads from actual storage systems and the simulation of replication selection algorithms, to understand the key features of the workload.

Required skills

The candidate needs a good level in algorithms and programming. Preliminary knowledge in scheduling and/or distributed systems is definitely a bonus.

Bibliography

[1] Vikas Jaiman, Sonia Ben Mokhtar, Vivien Quéma, Lydia Y. Chen, and Etienne Rivière. Héron: Taming tail latencies in key value stores under heterogeneous workloads. In *Proceedings of the International Symposium on Reliable Distributed Systems (SRDS)*, 2018.

[2] Arnaud Legrand, Alan Su, and Frédéric Vivien. Minimizing the stretch when scheduling flows of divisible requests. *J. Scheduling*, 11(5):381–404, 2008.

[3] Barbara Simons. Multiprocessor scheduling of unit-time jobs with arbitrary release times and deadlines. *SIAM Journal on Computing*, 12(2):294–299, 1983.