REGION
BOURGOGNE
FRANCHE
COMTE

EIPHI-BFC
GRADUATE SCHOOL
CROSS DISCIPLINARY SCIENCE
AND TECHNOLOGY

PhD Thesis Subject:
Stochastic Scheduling for HPC Systems

FEMTO-ST (Besançon), LIP (Lyon)

Grant : Region Bourgogne Franche-Comté and Graduate School EIPHI

PhD supervisors: Louis-Claude Canon (FEMTO-ST), Loris Marchal (LIP)
Contact: louis-claude.canon@univ-fcomte.fr

1 Scientific Context

Scheduling task executions in a parallel environment has been the focus of significant attention since the dawn of computer systems. A strong body of literature provides long-standing results such as the (2-1/m) performance bound for list heuristics (Graham, 1966). The research in this area has had a significant impact in HPC (High Performance Computing), which has lead to improvements in resources usage efficiency, power minimization, etc. By its common usage in optimizing everyday systems, scheduling represents a strong societal challenge.

Scheduling a set of applications on distributed resources constitutes a complex problem. The objective consists in defining the allocation of applications to machines to obtain the best possible performance. Numerous studies propose new algorithms and validate them by measuring the gain over existing solutions. To realise these measurements, we often rely on simulations that require to implement the scheduling algorithm. An application and platform model is also necessary to design relevant instances on which the scheduling algorithms are tested. In the proposed problem, the application is modeled by an execution cost for each task that depends on the executing machine. The way these instances are generated for the simulation can impact significantly the results. Beyond this general evaluation approach, it is relevant to focus on a preliminary analysis that characterizes directly the performance of the algorithms under study.

2 Objectives and Expected Results

The whole methodology consists in going back and forth between stochastic analysis and instance characterization by focusing on three questions as their interplay can lead to revisiting other questions when one of them progresses:

- stochastic analysis of algorithm quality and complexity;
- identification of instance distributions, which requires to identify their properties;
- analysis of existing empirical instances in actual HPC systems and applications.

The first axis will consists in analyzing the average performance of classical scheduling strategies (FIFO, EFT, LPT, SRPT, EDF) using specific instances. Many contributions have been proposed to optimize the expectation of the completion time for instance (Bruno et al., 1974, Han et al., 2018). However, these studies consider stochastic problem inputs, i.e., the data is not deterministically known. In our case, we assume that each input is known without uncertainties, and we are interested in the average performance over all inputs. Recent developments in the field of average-case analysis (Bogdanov et al., 2006) will help extending the scheduling field to more realistic distributions, which represents a crucial challenge to achieve our goals.

The second axis aims at characterizing the properties of the instances and their distribution. For instance, given that uniform task graphs may be easy to tackle, a question arises on how should the task graphs be distributed, if not uniformly. Designing generation methods, constrained on the distribution of the instances, completes this axis.

Finally, the third axis involves more empirical tools to study existing systems. Both the hardware and software can be part of an instance: while a task graph solely concerns the application, an execution time is the result of the interaction of both the application and the system. To the best of our knowledge, no data set has been analyzed in the HPC field with respect to the stochastic analysis of scheduling strategies and the generation of new instances.

Overall, the objective consists in focusing on classical strategies (FIFO, EFT, LPT, SRPT, EDF) and practical scheduling problem. In particular, the scheduling problem $P|r_j|F_{\max}$ is relevant to optimize key-value stores such as Apache Cassandra or MongoDV that are at the center of many Big Data systems. This problem consists in distributing replicated values on multiple machines and then allocate each request on these values to the best machine. Optimizing the maximum service time may prevent one of the most important problem faced by these systems: the tail latency, which is the delay incurred by the slowest requests. This problem involves a distribution of costs and an arrival model for the requests.

3 Planning and Localisation

The PhD work will take place between Besançon (Femto-ST) and Lyon (ENS Lyon) and will be supervised by Louis-Claude Canon, assistant professor at the University of Besançon, and Loris Marchal, CNRS researcher in team ROMA of LIP at ENS Lyon The PhD candidate will thus be located on these two sites. In particular, he will stay at least 12 months on each of these laboratories.

Expected start: September 2020.

4 Student Profile

The candidate will own a master degree or an equivalent diploma. The proposition is also open to students that do not have a diploma in computer science but with good background in optimization, probability and advanced skills in programming. The following qualities will be appreciated:

- Scientific curiosity
- Team work
- Advanced Skills in Programming
- Advanced Skills in Mathematical Tools for Computer Science

- Optimization and Probability
- Written and oral English

5 References

- Bruno, James, Edward G. Coffman Jr, and Ravi Sethi. "Scheduling independent tasks to reduce mean finishing time." Communications of the ACM 17.7 (1974): 382-387.
- Bogdanov, Andrej, and Luca Trevisan. "Average-case complexity." Foundations and Trends in Theoretical Computer Science 2.1 (2006): 1-106.
- Coffman, E. G., et al. "Probabilistic analysis of packing and related partitioning problems." Statistical Science 8.1 (1993): 40-47.
- Cole, Richard, and David C. Kandathil. "The average case analysis of partition sorts." European Symposium on Algorithms. Springer, Berlin, Heidelberg, 2004.
- P. Diaconis and L. S. Coste, “Random walk on contingency tables with mixed row and column sums,” Harvard University, Department of Mathematics, Tech. Rep., 1995.
- R.L. Graham. Bounds for certain multiprocessing timing anomalies. Bell System Technical Journal, 45(2):1563–1581, 1966.
- R.L. Graham. Bounds on multiprocessing timing anomalies. SIAM Journal on Applied Mathematics, 17(2):416–429, 1969.