Open Science and Reproducible Research on Distributed Systems

Post-doc position

Executive summary:
The goal of this post-doc position is to contribute to the opening of research on distributed systems by solving the challenges that must be overcome in order to allow Open Science and Reproducible Research in this field – improving description and publication of experiments and their results, facilitating the analysis of results, etc.

Key skills needed:
Mandatory:
- former experience with experimentation, preferably in the context of distributed systems (Cloud, HPC, P2P, Grids)
- performance evaluation, basic statistics

Appreciated:
- programming with high level languages (scripting languages such as Ruby)
- software development practices and tools, preferably in a Linux environment

Research team name: AlGorille (leaders: Jens Gustedt and Martin Quinson)
Location: LORIA, Nancy, France
Contact: Lucas Nussbaum <lucas.nussbaum@loria.fr>
Keywords: experimentation, distributed systems, open science, reproducible research

Context

Distributed systems research

Distributed systems such as grids, clusters, peer-to-peer systems, high-performance supercomputers, cloud computing infrastructures or desktop computing environments, benefit of an ever increasing popularity nowadays. Distributed applications (such as decentralized data sharing solutions, games, scientific application, high-traffic web applications or scientific computations) are executed routinely on these systems.

By nature, the resulting environments and applications are extremely complex and dynamic because they aggregate thousands of elements that are heterogeneous and shared among several users. This make these systems very challenging to study, test, and evaluate. Computer scientists traditionally study their systems a priori by reasoning theoretically on the constituents and their interactions. But the complexity of these systems make this methodology is near to impossible, explaining that most of the studies are done a posteriori through experiments.

Experimentation in distributed systems research
Three main methodologies exist to experiment with computer systems: real-scale, simulation and emulation. **Real-scale** (or *in situ*) consists in executing the real application under study on an experimental platform like Grid'5000 (a large scale experimental platform in France, composed of more than 1600 machines). On the opposite, with **simulation**, both the application and the environments are replaced by models, and the interactions between both models are computed by a simulator. **Emulation** can be seen as an intermediate approach where the real application is executed within a synthetic environment. Typically, one will use a homogeneous cluster of machines as an execution environment, and use an emulation layer to reproduce the complex conditions found on the real Internet.

The AlGorille team is deeply involved in all those methodologies. It has a leadership role in the world leading **Grid'5000** testbed and **SimGrid** simulator, and develops entirely the **Distem** emulator. We see those different methodologies as complementary approaches to work on the different steps of the scientific workflow: ideas are first maturated into algorithms in the simulator, before they are implemented in as prototypes tested on a real-scale testbed. Once the prototype is known to work, an emulator is used to evaluate the behavior of the prototype under various experimental conditions.

**Open Science and Reproducible Research**

The Open Science movement, which emerged in computational sciences, aims at conducting research in the spirit of free and open source software. It encompasses the publicity of software, data and results in order to allow enabled the detailed understanding of the experimental processes and obtained results. It also favors the reproduction of experiments and results, and the general increase of the quality of experiments.

**Description**

The goal of this post-doc position is to contribute to the opening of research on distributed systems – to solve the challenges that must be overcome in order to allow Open Science and Reproducible Research in this field. This will be done by focusing on the experimentation methodologies and tools actively worked on in the AlGorille team, but the designed solutions will aim at being more generic.

The central challenge in this topic is the description and publication of experiments (including parameters, environments, platforms, testbeds, data) and their results, in order to support provenance (traceability of results). This publication must ideally be achieved in formats that can be supported by several tools, in order to ease the comparison of results. Therefore, this work requires a deep understand of the current and future needs of simulators, emulators and testbeds. Some solutions have been designed in the context of other sciences, adapting those to our field require deep changes both in the solutions themselves, and in our experimental tools and testbeds.

As many experiments in our field have low statistical validity, another challenge is the adaptation and integration of Design of Experiments solutions into our experimentation frameworks, in order to facilitate the efficient execution of numerous experiments and their statistical analysis.

This work will be done in collaboration with other teams in order to be applied to real experiments campaigns, by leveraging AlGorille’s participation in the INRIA large-scale action Hemera ([https://www.grid5000.fr/mediawiki/index.php/Hemera](https://www.grid5000.fr/mediawiki/index.php/Hemera)) that involves 25 research teams in France.
Links and bibliography

- AlGorille research team: http://www.loria.fr/equipes/algorille/
- Grid'5000 testbed, https://www.grid5000.fr/
- SimGrid, http://simgrid.gforge.inria.fr/
- Distem, http://distem.gforge.inria.fr/
- Realis’2013 (satellite event of the ComPAS conference focusing on reproducibility of experiments)
  - website: http://compas2013.inrialpes.fr/evenements-satellites/realis

To apply, send a CV and a motivation letter to lucas.nussbaum@loria.fr