High performance serverless computing

High performance serverless computing

Keywords : runtime system, serverless computing, parallel programming

Context

In order to exploit a supercomputer, developpers reserve a set of server for a fixed period, eg 10 servers for 2 hours. Then, they deploy an MPI application on the reserved servers. The parallelism of the application may not be constant: some parts of the application are highly parallel and could run on hundreds of servers, while some other parts of the application are sequencial and only exploit one server.

Goal

The goal of this project is to design a runtime system that mixes HPC task programming (such as StarPU), and serverless computing (such as OpenWhisk). This would allow developpers to write parallel applications using task parallelism, and the runtime system would adapt number of servers to the application. This way, the highly parallel parts of the application would run on hundreds of servers, while the sequential part would only run on one machine. This could decrease the time to solution of applications while increasing the overal usage of supercomputers.

The main steps of this project are:

- study task programming runtime systems (StarPU, PaRSEC, ...), and serverless systems (eg. OpenWhisk, rFaaS, ...)
- design a runtime system for creating tasks, and executing them on multiple machines.
- evaluate the runtime system on a set of applications

Contact

François Trahay francois.trahay@telecom-sudparis.eu
Inria Benagil
Télécom SudParis