Intership offer for Master Student - Computer Science

Laboratory / University: IRIT / Université Paul Sabatier

Research teams: PYRAMID & SEPIA

Title: Development and deployment of data replication strategies on Grid5000

Recent infrastructure like Clouds consider specific characteristics as resource elasticity while taking into account cloud providers' businesses. This leads to the establishment of an economic model: the pay-as-you-go model. It means that the tenant pays only what it consumes as resources. The Service Level Agreement (SLA), a contract signed between the provider and the tenant, has also to be considered. This contract sets up Service Level Objective (SLO) the provider has to fulfill otherwise it will have to pay penalties to the concerned tenant. Among these objectives, we can cite availability and performance objectives. Furthermore, ecological considerations are getting more and more noticeable with an increasing impact of energy consumption reduction policies and so greenhouse gas reduction policies.

Data replication is a widely used technique upon distributed systems. It aims to increase data availability, reduce bandwidth consumption when accessing data and achieve fault-tolerance. Numerous data replication strategies have been proposed upon different architectures while taking into account characteristics of each one. On cloud architecture, these strategies should consider an elastic management of resources while satisfying data availability and performance objectives. Nowadays, satisfying other objectives like reducing the provider's expenditure or reducing energy consumption remains an interesting challenge to face of.

Grid5000 platform is a nation-wide testbed platform with more than 800 nodes distributed among 8 sites. This platform permits experiments on large-scale architectures. Many software tools are implemented in order to emulate nodes from different cities. They also permit estimating energy consumption of software and dockers on a set of nodes.

The goal of this internship is to develop and deploy multiple data replication strategies on Grid5000 nodes in order to compare them. These data replication strategies will be implemented on a distributed file management system such as Hadoop. Afterward, queries will be submitted from different nodes in order to access data such as NoSQL type data. Different workloads will be considered in order to carry out real experiments on physical infrastructures. Finally, this internship will take place at the IRIT Laboratory (Institut de Recherche en Informatique de Toulouse). It will be in support of a 3rd year PhD student.

References

[ATC17] Muhannad Alghamdi, Bin Tang, and Yutian Chen. Profit-based file replication in data intensive cloud data centers. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–7, Paris, France, May 2017. IEEE.

[DWF16] Miyuru Dayarathna, Yonggang Wen, and Rui Fan. Data Center Energy Consumption Modeling: A Survey. *IEEE Communications Surveys & Tutorials*, 18(1):732–794, 2016.

[SMP19] Morgan Séguéla, Riad Mokadem, and Jean-Marc Pierson. Comparing energy-aware vs. cost-aware data replication strategy. In *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, pages 1–8, October 2019.

Keywords: Cloud, Data replication, NoSQL, Deployment, Grid'5000

Required skills: Programming (Java, Python ou C), Tool deployment

Wage: 564€/mois **Location**: IRIT, Paul Sabatier University **Duration**: 5 or 6 months

Contact: Morgan Séguéla, morgan.seguela@irit.fr, Riad Mokadem (riad.mokadem@irit.fr) et Jean-Marc Pierson (jean-marc.pierson@irit.fr).