Offre de stage Master - Informatique

Laboratoire/ Université : IRIT/ Université Paul Sabatier Toulouse III

Équipes de recherche : Pyramide & SEPIA

SUJET : Développement et déploiement de stratégies de réplication de données sur Grid5000

Les infrastructures récentes telles que le Cloud se doivent de considérer une gestion élastique des ressources tout en prenant en compte l'aspect commercial pour les fournisseurs de Cloud public. Cela entraîne la mise en place d'un modèle économique 'Pay-as-you-go' qui signifie que l'utilisateur paie uniquement ce qu'il consomme comme ressources. Le Service Level Agreement (SLA), un contrat signé entre le fournisseur et le locataire, doit également être respecté. Coté locataire, ce contrat précise le montant payé par ce dernier au fournisseur pour la location des services. Dans le SLA, on retrouve également les objectifs de niveau de service que le fournisseur doit satisfaire au risque de payer des pénalités au locataire concerné. Parmi ces objectifs, on citera les objectifs de disponibilité et de performances. De plus, les considérations environnementales sont de plus en plus présentes dans l'esprit collectif augmentant ainsi l'impact de politiques réduisant la consommation énergétique et donc de la production de gaz à effet de serre.

La réplication de données, une technique largement utilisée dans les systèmes distribués, permet d'améliorer la disponibilité de données et de réduire le temps de réponses lors de l'accès à ces données. De nombreuses stratégies de réplication de données ont été proposées dans différentes architectures en tenant compte des spécificités de chacune de ces architectures. Dans les architectures Cloud, ces stratégies s'appuient sur l'élasticité pour le partage de ressources entre les différents locataires tout en satisfaisant les objectifs attendus par ces locataires, en termes de performances par exemple. De nos jours, la satisfaction d'autres objectifs tels que la réduction des dépenses du fournisseur ou encore de la consommation énergétique constituent un challenge intéressant à relever.

La plate-forme Grid5000 est une plate-forme d'expérimentation nationale présente sur 8 sites différents et contenant plus de 800 nœuds. Cette plate-forme permet de réaliser des expériences sur des architectures à large échelle. De plus, de nombreux outils sont mis en place sur cette plate-forme pour émuler des nœuds présents dans différentes villes. Ils permettent également d'estimer la consommation en puissance des logiciels et conteneurs sur plusieurs nœuds.

L'objectif de ce stage est de développer et de déployer plusieurs stratégies de réplications de données sur des noeuds de Grid5000 puis, de les comparer. Ces stratégies de réplication de données seront mises en place sur un système de gestion de fichiers distribués de type Hadoop. Par la suite, des requêtes seront mises en place afin d'interroger des bases de données de type NoSQL. Différentes charges de travail seront également considérées afin de réaliser des expérimentations réelles sur des infrastructures physiques. Enfin, ce stage se déroulera à l'IRIT (Institut de Recherche en Informatique de Toulouse) et se fera en soutient d'un doctorant en 3ème année de Thèse .

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.

Mots Clés : Cloud, Réplication de données, NoSQL, Déploiement, Grid'5000

Compétences attendues : Programmation (Java, Python ou C), Déploiement d'outils

Rémunération : 564€/mois **Lieu du stage** : IRIT, Université Paul Sabatier **Durée** : 5 ou 6 mois

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

Intership offer for Master Student - Computer Science

Laboratory/ University: IRIT / Université Paul Sabatier

Research teams: Pyramide & 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, NoSQL, Deployment, Grid'5000

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

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

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