Proposition de recherche doctorale

Software Defined Storage for Data Intensive Scalable Computing

Mots clés :

- **Directeur de thèse** : pietro MICHIARDI
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire de recherche d'EURECOM
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Abstract. The objective of this Thesis is to design and create a software-defined storage layer for data-intensive scalable computing applications, leveraging the Apache OpenStack platform. Such software-defined layer will enable the efficient execution of virtualized analytics applications over virtualized storage resources thanks to flexible, automated, and low cost data management models based on software-defined storage (SDS). In order to achieve this objective, the Thesis will focus on the following topics: • Storage and compute disaggregation and virtualization. Virtualizing data analytics to reduce costs implies disaggregation of existing hardware resources. This requires the creation of a virtual model for compute, storage and networking components that allows orchestration tools to manage resources in an efficient manner. For the orchestration layer it is essential to define and implement innovative scheduling mechanisms so that the provisioning of virtual components for the analytics platform is made to achieve performance guarantees; • SDS Services for Analytics. An important objective of the Thesis is to define, design, and build a software stack of SDS data services enabling virtualized analytics with improved performance and usability. Among these services, we will consider native object store analytics that will allow running analytics close to the data without taxing initial migration. Data reduction services that will be optimized for the special requirements posed by virtualized analytics platforms, and specialized persistent caching mechanisms, advanced prefetching, and data placement will complement the resource allocation components.

Résumé du projet de recherche (Langue 2)

• Orchestration and deployment of “Big Data” analytics services. The ultimate goal of the Thesis is to design and build efficient deployment strategies for virtualized analytic-as-a-service instances, which take form of novel scheduling mechanisms. In particular, the focus of this work is on data-intensive scalable computing (DISC) systems such as Apache Hadoop and Apache Spark, which enable users to define both batch and latency-sensitive analytics. This objective includes the design of scalable algorithms that strive at optimizing a service-wide objective function (e.g., optimize performance, minimize cost, etc ...) under heterogeneous workloads. Ultimately, the candidate will create an experimental prototype of the software-defined storage layer and the scheduling components outlined above, on top of OpenStack. To this end, the Candidate will contribute to widely used OpenStack projects including OpenStack Swift, OpenStack Nova, OpenStack Cinder and OpenStack Sahara.