**Proposition de recherche doctorale**

**Design, Analysis and Implementation of Large-scale Distributed Systems for Data-intensive Applications**

**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)**

The objective of this Thesis is to design, analyze and experiment with tools for data-intensive applications. The candidate will begin with an experimental work towards the definition of scalable algorithms to analyze large-scale network traces using MapReduce, a recent system designed to perform parallel processing of massing amounts of data. The design and analysis of scalable (parallel) algorithms is not trivial as it involves a deep understanding of the underlying execution framework, especially for what concerns the optimization of resource utilization (namely, disk and network I/O): as such, the candidate will be required to “bend” the MapReduce framework in order to accommodate the execution flow of the algorithms. In addition to studying the programming model, the candidate will focus on the key building blocks of the MapReduce framework as well. In particular, the candidate will work on scheduling mechanisms that enable sharing the resources of a cluster of commodity machines for multiple concurrent jobs. Insights from the theory of scheduling will have to be put into practice with the goal of producing a scheduling protocol that caters fairness and performance. This work requires the candidate to participate and contribute to the open-source community behind Hadoop, a Java-based implementation of MapReduce, by issuing a JIRA (a ticket) and by launching a series of experiments to evaluate and compare the schedule of several other protocols. As a broader research direction, in this Thesis the candidate will cover the measurement and analysis of the impact of cluster virtualization on parallel processing frameworks, including MapReduce. Eventually, the output of this task will be the design and evaluation of (virtual) resource allocation mechanisms to deploy virtual instances of cluster machines on a physical cluster, while taking into account the requirements of the parallel processing framework (e.g. data locality).

**Résumé du projet de recherche (Langue 2)**

The objective of this Thesis is to design, analyze and experiment with tools for data-intensive applications. The candidate will begin with an experimental work towards the definition of scalable algorithms to analyze large-scale network traces using MapReduce, a recent system designed to perform parallel processing of massing amounts of data. The design and analysis of scalable (parallel) algorithms is not trivial as it involves a deep understanding of the underlying execution framework, especially for what concerns the optimization of resource utilization (namely, disk and network I/O): as such, the candidate will be required to “bend” the MapReduce framework in order to accommodate the execution flow of the algorithms. In addition to studying the programming model, the candidate will focus on the key building blocks of the MapReduce framework as well. In particular, the candidate will work on scheduling mechanisms that enable sharing the resources of a cluster of commodity machines for multiple concurrent jobs. Insights from the theory of scheduling will have to be put into practice with the goal of producing a scheduling protocol that caters fairness and performance. This work requires the candidate to participate and contribute to the open-source community behind Hadoop, a Java-based implementation of MapReduce, by issuing a JIRA (a ticket) and by launching a series of experiments to evaluate and compare the schedule of several other protocols. As a broader research direction, in this Thesis the candidate will cover the measurement and analysis of the impact of cluster virtualization on parallel processing frameworks, including MapReduce. Eventually, the output of this task will be the design and evaluation of (virtual) resource allocation mechanisms to deploy virtual instances of cluster machines on a physical cluster, while taking into account the requirements of the parallel processing framework (e.g. data locality).