Offre de stage : Traitement performant des données de CTA (Cherenkov Telescope Array)

(Stage sous réserve de financement)

Encadrants : Luisa Arrabito (arrabito@in2p3.fr)

Sujet :
CTA, Cherenkov Telescope Array [1], est une collaboration internationale regroupant environ 1200 membres venant d'une centaine d'instituts de recherche. L'équipe de production de CTA, basée au Laboratoire Univers et Particules de Montpellier (LUPM [2], CNRS), est en charge de la production des simulations Monte Carlo, dont l'objectif est de caractériser la réponse des réseaux des télescopes. Ces simulations sont obtenues au travers un ensemble de 'productions', où chaque production consiste à exécuter des centaines de milliers de tâches de calcul sur la grille européenne (European Grid Infrastructure, EGI). Chaque campagne de simulation génère plusieurs centaines de milliers de fichiers, pour un volume correspondant de l'ordre de quelques centaines de TB, également archivés sur la grille. Afin de gérer ces productions massives, l'équipe de production utilise et contribue au logiciel DIRAC [3]. DIRAC est un projet 'open source' (sous licence GPL V3, utilisant GitHub) pour la gestion de calculs et de données dans des environnements distribués. Conçu avec une architecture 'service-oriented', DIRAC emploie des 'agents' qui interrogent régulièrement des bases de données spécifiques, afin de construire et soumettre les différentes tâches de calcul. Toutefois, ce mode de fonctionnement montre des limitations en termes de performances.
L'objectif du stage est d'explorer et d'implémenter un nouveau paradigme pour la construction des tâches de calcul, qui serait plus performant du mode actuel. Un défi majeur consiste dans le développement d'un système évolutif, capable de gérer le grand nombre de fichiers traités par plusieurs productions concurrentes. Une autre difficulté réside dans le fait d'assurer que la totalité des fichiers soit traitée.
Le stagiaire conduira toutes les phases du projet, en partant de l'analyse des besoins jusqu'à la vérification de l'implémentation, en suivant un processus de certification établi. Enfin, le logiciel sera mis en production et appliqué aux simulations de CTA. L'étudiant aura l'opportunité de travailler dans une équipe internationale de chercheurs et ingénieurs. Du point de vue technique, il approfondira ses connaissances en développement python, tout en découvrant les problématiques liées au calcul distribué.

Compétences requises :
Un niveau de compétence intermédiaire en python est requis. Des connaissances relatives à au moins un système de Message Queuing sont un atout.

Liens utiles :
[1] https://portal.cta-observatory.org/Pages/Home.aspx
[2] http://www.lupm.univ-montp2.fr/
[3] http://diracgrid.org/

Title: Efficient processing of large scale CTA (Cherenkov Telescope Array) data

Supervisors: Luisa Arrabito (arrabito@in2p3.fr)

(Internship subject to the funding agreement)

Content:
CTA, Cherenkov Telescope Array [1], is a worldwide collaboration gathering about 1200 scientists from a hundred of institutes. The CTA production team, based at Laboratoire Univers et Particules de Montpellier (LUPM [2], CNRS), regularly performs massive Monte Carlo simulations aimed to characterize the instrument

response. These simulations are obtained through a set of ‘productions’, where each production consists in the execution of hundreds of thousands of similar computing tasks on the European Grid Infrastructure (EGI). The results of each simulation campaign are also archived on the Grid and they are of the order of hundreds of TB and millions of files. In order to manage such large productions, the production team uses and contributes to a software system called DIRAC [3]. DIRAC is an open source project (under GPL V3 licence and using GitHub), following a service-oriented paradigm, for the workload and data management in distributed environments. In particular, in order to build and submit the various computing tasks, DIRAC employs a number of agents periodically querying specific databases. However, this mode has some performance limitations.

The aim of the internship is to explore and implement a new paradigm for the construction of the computing tasks, which would be more efficient than the current one. One major challenge consists in developing a scalable system able to manage the large number of files to be treated by several productions running in parallel. Another difficulty is to ensure that no single file remains unprocessed or partially processed.

The student will lead all the phases of the project, starting from analysing the requirements to the verification of the implementation, following an established certification process. Finally, the software will be released and applied to real use-cases coming from CTA. The student will have the opportunity to work in an international team of scientists and engineers. From the technical point of view, he will develop python coding skills and will discover the problematic related to the distributed computing.

Requisites:

The student should have intermediate knowledge of the python programming language. Knowledge of at least one Message Queuing system is a plus.

Useful links:

[1] https://portal.cta-observatory.org/Pages/Home.aspx
[2] http://www.lupm.univ-montp2.fr/
[3] http://diracgrid.org/