Analyse d'une trace de données de Google
ING3-IMSI – Projet IMSI
Année 2016–2017

1 Introduction

La trace d'utilisation d'un cluster de Google [1] est un ensemble de données collecté par Google sur l'un de leurs clusters pendant le mois de mai de 2011. Contenant une quantité considérable d'information (41 GBytes), elle a été mise à disposition de la communauté scientifique afin d'encourager la recherche sur l'analyse des données des grands centres de calcul.

2 Ressources

La trace est disponible sur le *github* de Google [2], mais une copie vous sera fournie par votre encadrant. La documentation du format et des spécifications de la trace sont aussi en ligne [3]. De plus, il existe un forum en Internet [4] où vous pouvez trouver des réponses à plusieurs questions. Finalement, il existe une considérable quantité d'articles scientifiques sur l'analyse de la trace [5].

À cause de sa taille, la trace ne peut être analysée confortablement qu'en utilisant un cluster de calcul dont vous recevrez un compte de connexion.

3 Modalités

Ce projet est à effectuer par groupe de 2, la composition des groupes est libre. Les choix algorithmiques et d'implémentation sont libres. Le code et les outils d'analyse sont libres. Il est recommandé d'utiliser le GitLab de l'école (codes sources, rapport, fichiers de tests, etc.).

4 Objectifs

Le but de ce projet est d'analyser et caractériser certains aspects de l'utilisation du cluster. Au moins, les questions suivantes doivent être traitées :

1. **Caractérisation statique de la charge :**
   - Identification des travaux et tâches (*jobs* et *tasks* dans la nomenclature de Google) les plus prenants en CPU, en valeurs maximales et en moyenne pendant la durée de la trace. Ces travaux seront appelés désormais *jobs dominants*.
   - Identification des jobs et tasks les plus prenants en mémoire vive, en valeurs maximales et en moyenne pendant la durée de la trace.
   - Classification des jobs dominants par classe de priorité.
   - Pour les jobs dominants, étude de la corrélation entre la consommation de CPU et de la mémoire vive.
2. **Caractérisation dynamique de la charge :**
   - Pour les jobs dominants, étude de sa périodicité classifiée par classe de priorité.

5 Cahier des charges

Vous devez établir votre cahier des charges pour la deuxième séance de projet (la veille à 23h59 au plus tard et devra également être publié sur Arel). Celui-ci devra contenir une reformulation du sujet faisant clairement apparaître votre compréhension du problème, la stratégie que vous allez adopter ainsi que le planning prévisionnel d'exécution. En plus, le cahier des charges devra contenir les informations suivantes :

- Votre compréhension sur votre renseignement initial du format de la trace.
- L'utilisation du cluster que vous avez prévu pour l'analyse de la trace.
- L'élection du langage et des outils d'analyse.
- Le nettoyage initial des données que vous envisagez nécessaire.

À la fin de la deuxième séance on fera le point pour discuter et réorienter (s'il y en a besoin) votre stratégie.

6 Déroulement des séances

Tout au long des séances, vous devrez rendre compte de votre travail à votre encadrant Le rendu final sera le dernier rendu publié sur GitLab, au plus tard à 23h59 le **27/02/2017**. Il sera également publié sur Arel.

7 Notation

Seront notées la vraisemblance de l'implémentation des algorithmes, la cohérence entre la réalisation et le cahier des charges, les techniques et méthodes d'implémentation, la gestion du suivi de projet et, bien entendu, la réponse apportée à l'objectif attendu.

Un rapport devra être fourni avec le rendu du code et être suffisamment pertinent, contenant notamment votre avis sur l'opportunité des méthodes employées. Il ne sera pas noté mais pourra pénaliser lourdement la note donnée à la réalisation s'il se trouvait ne pas être satisfaisant.

Références

[1] J. Wilkes, *More Google cluster data, 2011, (Google research blog).*, 2011. Posted at `http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html`.

[2] J. Wilkes and C. Reiss, *ClusterData2011_2 traces*, 2011. `https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md`.

[3] J. H. C. Reiss, J. Wilkes, *Google cluster-usage traces: format + schema, Technical Report, Google Inc., Mountain View, CA, USA, 2011 . Revised 2012.03.20.* Posted at `https://drive.google.com/open?id=0B5g07T_gRDg9Z0lsSTEtTWtpOW8&authuser=0`.

[4] *Google cluster data - discussions.* `https://groups.google.com/forum/#!forum/googleclusterdata-discuss`.

[5] *Google cluster traces bibliography.* `https://github.com/google/cluster-data/blob/master/bibliography.bib`.