

Allocation de ressources par une méthode hybride machine learning - optimisation en contexte de conteneurs

Etienne Leclercq ^{*,**} Jonathan Rivalan ^{*}, Frédéric Roupin ^{**}, Céline Rouveirol ^{**}

^{*}Alter way - (etienne.leclercq,jonathan.rivalan)@alterway.fr,

^{**}LIPN - (leclercq, roupin, rouveirol)@lipn.univ-paris13.fr

1 Introduction

La gestion des ressources est l'un des défis les plus importants du cloud et est composé principalement du problème de placement des services sur l'infrastructure, défi difficile notamment par l'introduction de l'élasticité. Ceci rend la modélisation du problème d'affectation des conteneurs complexe, car cette élasticité entraîne de l'inconnu avant l'affectation. Lorsque l'on cherche une bonne solution pour le placement des conteneurs, l'objectif global est de limiter la surcharge des serveurs qui se traduit par des interruptions de services, et d'optimiser l'utilisation globale des ressources, comme le nombre de serveurs en cours d'utilisation. Pour atteindre ces objectifs, nous considérons 4 critères qui représentent le nombre de nœuds, l'amplitude et la variance de la consommation des nœuds, et le nombre de dépassement de capacité des nœuds. En pratique, le scheduler de Docker ne tient pas compte de l'utilisation des ressources pour placer les conteneurs dans un cluster (Hassen, 2016). D'autres travaux se sont intéressés au placement de conteneurs (Tosatto et al., 2015), mais sans possibilité de l'ajuster en fonction de l'évolution des consommations. L'ajustement des ressources allouées aux conteneurs a également été étudié, mais sans modifier leur placement sur l'infrastructure. Pour répondre à cette problématique et les objectifs cités, nous avons développé une méthode alliant apprentissage et optimisation. Cette interaction constitue par ailleurs un intérêt supplémentaire, étant encore peu développée aujourd'hui, un domaine étant souvent utilisé pour améliorer les méthodes de l'autre. Ici, le clustering est utilisé pour identifier les profils de consommation et guider le placement en utilisant des techniques d'optimisation.

2 Méthodologie - la boucle de rétro-action

Comme en production, de nouvelles données sont collectées en continu, nous avons développé un mécanisme de *boucles de rétro-action*, qui vise à évaluer la solution actuelle avec de nouvelles données, et à mettre à jour cette solution si besoin. Par ailleurs, les services fonctionnant en temps réel, la solution doit être trouvée rapidement, la modélisation du problème complet avec tous ses critères et contraintes ne peut être envisagée. Ainsi nous utilisons des informations de clustering, mises à jour dans le temps grâce à des méthodes d'optimisation, pour évaluer la solution actuelle d'affectation et la mettre à jour si besoin. Les problèmes de clustering et de placement sont donc modélisés par des problèmes d'optimisation, dont les relaxations continues sont considérées. A ces problèmes nous ajoutons des contraintes de type *mustLink* pour indiquer les solutions actuelles, où $u_{i,j}$ et $v_{i,j}$ sont des variables spécifiant si i