PER2023–052 - Type: recherche

Limiting Memory Reclaiming Impact on VMs Performance

Etudiant : Tobias Bonifay (SI 5 – AL)

Encadrants : Dino Lopez Pacheco (I3S), Ramon APARICIO PARDO (I3S)

1. Résumé exécutif

Notre projet vise à développer une méthode avancée pour optimiser la gestion de la mémoire dans les environnements de virtualisation, réduisant ainsi l'impact négatif sur les performances des machines virtuelles (VM). En s'appuyant sur les fondations de la virtualisation comme moteur d'efficacité dans les data centers modernes, nous nous concentrons sur l'amélioration des techniques de récupération de mémoire - une ressource souvent limitée en raison du grand nombre de VM requises. Les techniques actuelles, telles que le "ballooning" et le "swapping", bien que fonctionnelles, peuvent mener à une dégradation des performances des VM, affectant la qualité de service.

Notre innovation réside dans l'application de l'apprentissage par renforcement (RL) et de l'apprentissage par imitation (IL) pour former des politiques de contrôle de la mémoire qui minimisent l'impact sur la performance sans intrusions lourdes dans les VM. En modélisant ce problème comme un processus décisionnel de Markov (MDP), nous cherchons à prédire de manière dynamique et à contrôler la quantité de mémoire à récupérer en fonction de l'activité des VM, en temps réel.

Nous anticipons que notre solution permettra une gestion plus efficace des ressources de mémoire, tout en préservant des performances optimales pour les applications gourmandes en ressources, comme les serveurs web et les bases de données en mémoire.

2. Description du projet

Contexte technologique

- Virtualisation des serveurs dans les centres de données modernes utilisant des hyperviseurs comme KVM et VMware pour optimiser l'utilisation des ressources physiques.
- Mécanismes de partage et de réclamation de la mémoire dans les environnements virtualisés, confrontant les techniques traditionnelles de gestion de la mémoire comme le swapping et le ballooning.
- Utilisation des techniques d'apprentissage automatique, spécifiquement l'apprentissage par renforcement (Reinforcement Learning - RL) et l'apprentissage par imitation (Imitation Learning - IL), pour optimiser la gestion des ressources.

Motivations

- Réduire l'impact négatif des techniques de réclamation de mémoire sur la performance des VMs, essentiel pour maintenir une haute qualité de service.
- Pousser l'efficacité des centres de données en maximisant l'utilisation de la mémoire physique sans ajouter d'équipement supplémentaire, ce qui est économiquement et écologiquement bénéfique.
- Progression scientifique dans l'application de modèles d'apprentissage automatique aux problématiques systèmes, un domaine encore peu exploré.

Objectifs à atteindre

- Développer une solution pour réclamer la mémoire inutilisée des VMs (machine-virtuelles) sans dégrader la qualité de service perçue par l'utilisateur.
- Tester et implémenter les solutions de réclamation de mémoire et les heuristiques d'estimation de l'ensemble de travail actif sur un hyperviseur KVM.
- Développement des solutions de surveillance non intrusives pour estimer le niveau de QoS offert par une VM
- Élaborer une approche RL/IL pour déterminer les politiques de contrôle optimales pour la réclamation de la mémoire.

Risques identifiés (et contremesures)

- Inexactitude des prédictions du modèle ML pouvant conduire à une sur ou sous-estimation des besoins en mémoire des VMs. Contremesure : Validation croisée des prédictions avec des données historiques et ajustement des modèles.
- Performances insuffisantes lors de la mise à l'échelle. Contremesure : Tests de charge progressifs et optimisation des ressources.

Scenarios

Scénario 1 - Gestion dynamique de la mémoire pour une VM sous-utilisée :

La VM réduit sa consommation de mémoire sans impact sur le temps de réponse des applications.

Scénario 2 - Adaptation en temps réel à un pic de demande :

En cas de forte demande, réallocation de la mémoire dynamiquement pour maintenir la qualité de service.

3. Mise en en œuvre

Activités déjà réalisées :

Étude des recherches existantes et compréhension des techniques de réclamation de mémoire.

Configuration initiale de l'environnement de test avec VMs.

Premiers pas vers le développement d'un modèle de machine learning pour prédire l'utilisation de la bande passante.

Activités prévues :

Semaine 1

Mesure des performances de base des VMs avec httperf.

Début de la collecte de données pour l'apprentissage automatique.

Semaine 2

Écriture (Ajout sur l'existant) d'un script Python pour la capture et l'analyse des paquets.

Pré-traitement des données collectées, premiers essais d'entraînement

Semaine 3

Ajustement du modèle.

En tant qu'unique membre de l'équipe, je gérerai l'ensemble des tâches.