République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Supérieure en Informatique -08 Mai 1945- Sidi Bel Abbès

THESIS

To obtain the diploma of Engineer

Field: Computer Science
Specialty: Information Systems and Web

MicroService Instances Load Balancing in Fog Computing Using DRL

Presented by: **Emziane Mohamed**
Submission Date: **September, 2024**

In front of the jury composed of:
Ms. BENCHERIF Kheira, President
Mr. MALKI Abdelhamid, Supervisor
Ms. LEHIRECHE Nesrine, Examiner

Academic Year: **2023/2024**

0.1 Abstract

Fog-native computing is an emerging paradigm that facilitates the development of flexible and scalable IoT applications through the use of microservice architecture at the network edge. This paradigm allows IoT users to request and consume services in close proximity, thereby enhancing Quality of Service (QoS) attributes such as reduced service delays. Typically, these requested services consist of multiple interdependent microservice instances, collectively referred to as a service plan. However, selecting the appropriate service plan for each request is a challenging task in dynamic fog environments, where network conditions and service requests fluctuate rapidly.

In this research, we investigate the microservice selection problem for IoT applications deployed on fog computing platforms and propose a learning-based solution employing deep reinforcement learning (DRL). We implemented the fog computing platform as a multinode Kubernetes cluster with Raspberry Pi edge nodes and a central control plane. Real-time cluster metrics were collected, allowing us to observe the impact of DRL on load balancing and resource optimization. Our approach aims to optimize application request delays while effectively balancing the load among microservice instances. By leveraging DRL, the system can adapt to the dynamic nature of fog environments, learning to compute optimal physical service plans that accommodate varying demands and conditions.

To address the plan-dependency issue, we implement the proposed learning algorithm using action masking. This approach restricts the DRL agent's actions in specific states, thus preventing the selection of invalid service plans. Action masking aim to enhance the efficiency and effectiveness of the agent in selecting suitable service plans, ensuring optimized performance and load balancing in fog-native computing environments.

Keywords: MicroService, Load Balancing, Middle-wares, Fog Computing, Kubernetes Multi-node Cluster, Deep Reinforcement Learning

0.2 Résumé

L'informatique native au brouillard est un paradigme émergent qui facilite le développement d'applications IoT flexibles et évolutives en utilisant l'architecture de microservices à la périphérie du réseau. Ce paradigme permet aux utilisateurs IoT de demander et de consommer des services à proximité, améliorant ainsi les attributs de Qualité de Service (QoS) tels que la réduction des délais de service. En général, ces services demandés sont composés de plusieurs instances de microservices interdépendantes, collectivement appelées un plan de service. Cependant, la sélection du plan de service approprié pour chaque demande est une tâche difficile dans les environnements de brouillard dynamique, où les conditions réseau et les demandes de service fluctuent rapidement.

Dans cette étude, nous examinons le problème de la sélection de microservices pour les applications IoT déployées sur des plateformes de calcul en brouillard et proposons une solution basée sur l'apprentissage utilisant le deep reinforcement learning (DRL). Nous avons mis en place la plateforme de fog computing sous forme d'un cluster Kubernetes multinœud avec des nœuds périphériques Raspberry Pi et un plan de contrôle central. Les métriques en temps réel du cluster ont été collectées à l'aide de Grafana et Prometheus, ce qui nous a permis d'observer l'impact du DRL sur l'équilibrage de charge et l'optimisation des ressources. Notre approche vise à optimiser les délais des demandes d'application tout en équilibrant efficacement la charge entre les instances de microservices. En exploitant le DRL, le système peut s'adapter à la nature dynamique des environnements de brouillard, apprenant à calculer des plans de service physique optimaux qui répondent aux demandes et conditions variées.

Pour remédier au problème de dépendance aux plans, nous mettons en œuvre l'algorithme d'apprentissage proposé en adoptant le masquage des actions. Cette approche limite les actions de l'agent d'apprentissage par renforcement profond (DRL) dans des états spécifiques, empêchant ainsi la sélection de plans de service invalides. Le masquage des actions vise à améliorer l'efficacité et l'efficience de l'agent dans la sélection de plans de service appropriés, garantissant ainsi des performances optimisées et un équilibrage de charge dans des environnements informatiques fog-natifs.

Mots-clés : Microservice, Équilibrage de Charge, Informatique en Brouillard, Cluster Kubernetes Multinœud Apprentissage par Renforcement Profond