الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbes

Master's Thesis

To obtain the diploma of **Master's Degree**

Field of Study: **Computer Science**

Specialization: **Computer Systems Engineering**

Theme

Artificial Intelligence Techniques for Load Balancing in Kubernetes-Orchestrated Microservice Architectures

---

Presented by
Harche Samir

Defended on: **15 September, 2025**
In front of the jury composed of

Dr. MALKI Abdelhamid	President of the Jury
Dr. Abdelhak-Djamel SERIAI	Thesis Supervisor
Dr. Bensenane Hamdan	Co-Supervisor
Dr. BEZZAOUCHA Fatima Souad	Examiner

Academic Year: 2024/2025

Abstract

The adoption of microservice architectures, particularly when orchestrated by platforms like Kubernetes, has become a standard for developing scalable and resilient applications. However, the dynamic and distributed nature of these systems introduces significant operational complexity, rendering traditional, rule-based load balancing techniques ineffective. This thesis presents a systematic review and a comprehensive state-of-the-art survey on the application of Artificial Intelligence (AI) to address the challenges of dynamic load balancing in these modern environments.

The study begins by establishing the foundational principles of microservices and critically analyzing the documented shortcomings of classical load balancing algorithms. It then provides a structured taxonomy of modern, intelligent approaches, systematically classifying methods from Machine Learning, Reinforcement Learning, and hybrid paradigms. The synthesis of the literature reveals a definitive trend towards adaptive, learning-based systems, with Reinforcement Learning emerging as a dominant paradigm for autonomous decision-making in complex Kubernetes environments.

The primary contribution of this work is a consolidated overview that maps the current research landscape, identifies key challenges such as model interpretability and multi-objective optimization, and highlights promising directions for future academic inquiry.

Keywords: Microservices, Load Balancing, Artificial Intelligence, Machine Learning, Reinforcement Learning, Kubernetes, Systematic Review.

Résumé

L'adoption des architectures microservices, en particulier lorsqu'elles sont orchestrées par des plateformes comme Kubernetes, est devenue une norme pour le développement d'applications scalables et résilientes. Cependant, la nature dynamique et distribuée de ces systèmes introduit une complexité opérationnelle significative, rendant les techniques traditionnelles d'équilibrage de charge basées sur des règles inefficaces. Cette thèse présente une revue systématique et une étude complète de l'état de l'art sur l'application de l'Intelligence Artificielle (IA) pour relever les défis de l'équilibrage de charge dynamique dans ces environnements modernes.

L'étude commence par établir les principes fondamentaux des microservices et par analyser de manière critique les lacunes documentées des algorithmes classiques d'équilibrage de charge. Elle fournit ensuite une taxonomie structurée des approches modernes et intelligentes, classifiant systématiquement les méthodes issues de l'Apprentissage Automatique (Machine Learning), de l'Apprentissage par Renforcement (Reinforcement Learning) et des paradigmes hybrides. La synthèse de la littérature révèle une tendance définitive vers des systèmes adaptatifs basés sur l'apprentissage, l'Apprentissage par Renforcement émergeant comme un paradigme dominant pour la prise de décision autonome dans les environnements Kubernetes complexes.

La contribution principale de ce travail est une vue d'ensemble consolidée qui cartographie le paysage actuel de la recherche, identifie les défis clés tels que l'interprétabilité des modèles et l'optimisation multi-objectifs, et met en évidence des pistes prometteuses pour de futures recherches universitaires.

Mots-clés: Microservices, Équilibrage de Charge, Intelligence Artificielle, Apprentissage Automatique, Apprentissage par Renforcement, Kubernetes, Revue Systématique.

ملخص

أصبح اعتماد معماريات الخدمات المصغرة، خاصة عند تنسيقها بواسطة منصات مثل Kubernetes، معيارًا لتطوير تطبيقات قابلة للتوسع ومرنة. ومع ذلك، فإن الطبيعة الديناميكية والموزعة لهذه الأنظمة تفرض تعقيدات تشغيلية كبيرة، مما يجعل تقنيات موازنة الحمل التقليدية القائمة على قواعد ثابتة غير فعالة. تقدم هذه الأطروحة مراجعة منهجية ومسحًا شاملًا لأحدث ما توصل إليه العلم في مجال تطبيق الذكاء الاصطناعي (AI) لمواجهة تحديات موازنة الحمل الديناميكية في هذه البيئات الحديثة.

تبدأ الدراسة بتأسيس المبادئ الأساسية للخدمات المصغرة وتحليل نقدي للعيوب الموثقة في خوارزميات موازنة الحمل الكلاسيكية. ثم تقدم تصنيفًا منظمًا للنهج الحديثة والذكية، حيث تصنف بشكل منهجي الأساليب المستمدة من التعلم الآلي (Machine Learning) والتعلم المعزز (Reinforcement Learning) والنماذج الهجينة. يكشف تجميع الأدبيات عن توجه واضح نحو الأنظمة التكيفية القائمة على التعلم، مع بروز التعلم المعزز كنموذج سائد لاتخاذ القرارات المستقلة في بيئات Kubernetes المعقدة.

تتمثل المساهمة الرئيسية لهذا العمل في تقديم نظرة عامة موحدة ترسم خريطة المشهد البحثي الحالي، وتحدد التحديات الرئيسية مثل قابلية تفسير النماذج والتحسين متعدد الأهداف، وتسلط الضوء على الاتجاهات الواعدة للبحث الأكاديمي المستقبلي.

الكلمات المفتاحية: الخدمات المصغرة، موازنة الحمل، الذكاء الاصطناعي، التعلم الآلي، التعلم المعزز، Kubernetes، مراجعة منهجية.