

Семенов С.Г.,
д.т.н., професор,
Єнгаличев С.О.,
здобувач третього (наукового) рівня вищої освіти,
Харківський національний економічний університет ім. Семена Кузнеця

АДАПТИВНЕ ПЛАНУВАННЯ ЗАДАЧ У РОЗПОДІЛЕНИХ СИСТЕМАХ НА ОСНОВІ АЛГОРИТМУ ОПТИМІЗАЦІЇ ПОЛІТИК ІЗ ПРОКСИМАЛЬНИМ ОНОВЛЕННЯМ

Проблема ефективного планування задач у гетерогенних розподілених інформаційних системах (HDCS) є однією з найбільш актуальних у сучасних обчислювальних технологіях. Висока динамічність середовища, різномірність ресурсів та ймовірнісна природа взаємозв'язків між задачами вимагають розробки нових підходів, що виходять за межі класичних статичних алгоритмів. Традиційні моделі планування, які базуються на детермінованих уявленнях про систему, не здатні повною мірою врахувати постійні зміни в навантаженні, обчислювальній потужності ресурсів і непередбачувані збої, що значно обмежує їхню ефективність у реальних розподілених середовищах [1].

Запропонований у роботі метод адаптивного планування задач базується на синергії трьох ключових компонентів: графових моделей типу DAG (Directed Acyclic Graph) і GERT (Graphical Evaluation and Review Technique), методів прогнозування на основі графових нейронних мереж (Graph Attention Network, GAT) і підкріплювального навчання з використанням алгоритму оптимізації політик із проксимальним оновленням (Proximal Policy Optimization, PPO). Поєднання цих технологій дозволяє не тільки точно прогнозувати критичний шлях задач у стохастичних умовах, а й здійснювати гнучке й адаптивне розподілення обчислювальних завдань залежно від актуального стану ресурсів системи.

Особливістю розробленого підходу є використання графової моделі DAG+GERT, яка дозволяє враховувати ймовірнісні переходи між задачами, що істотно підвищує стійкість планування до змін у середовищі виконання. Для оцінки критичного шляху застосовується Graph Attention Network, яка, завдяки механізму самопридільної уваги, визначає пріоритетні задачі й адаптується до

змін у структурі залежностей між підзадачами [2]. Це дозволяє значно точніше визначати ключові вузли, які впливають на загальний час завершення процесу, і оптимізувати ресурсні витрати.

Адаптивне керування задачами реалізується через агента PPO, який вчиться приймати рішення щодо розподілення задач у реальному часі, виходячи з інформації про поточний стан системи, прогнозоване навантаження та можливі ризики перевантаження або відмови окремих ресурсів. Завдяки цьому забезпечується мінімізація середнього часу виконання задач (makespan), вирівнювання навантаження між вузлами та максимізація стійкості до змін доступності обчислювальних потужностей [3]. Під час навчання агенту PPO забезпечується контроль стабільності оновлення політик, що запобігає їх нестабільній зміні під впливом локальних флуктуацій у даних.

Окремо впроваджено механізми оптимізації призначення задач на обчислювальні вузли за допомогою методів байєсівської оптимізації, що дозволяє враховувати не тільки поточне завантаження ресурсів, але й прогнозовані зміни у їх доступності. Це робить розподіл задач ще більш точним та ефективним у динамічному середовищі, яке притаманне сучасним хмарним платформам, великим кластерам і IoT-мережам.

Результати чисельного моделювання підтвердили, що використання запропонованого підходу дозволяє зменшити середній час виконання задач на 31,7 %, знизити дисбаланс навантаження між вузлами майже вдвічі порівняно з базовими алгоритмами FCFS, HEFT та класичними DAG-підходами. Аналіз експериментальних даних продемонстрував здатність моделі адаптуватися до раптових змін навантаження, появи нових задач у системі або непередбачених збоїв вузлів.

Практичне значення роботи полягає у можливості застосування розробленого методу для покращення ефективності роботи хмарних платформ, обробки великих обсягів даних у реальному часі, балансування навантаження у мультиагентних системах, а також для задач автоматичного управління в системах Інтернету речей (IoT) і в промислових кіберфізичних системах.

Гнучкість і масштабованість запропонованої архітектури робить її придатною для подальшого розвитку та інтеграції із сучасними фреймворками обчислювальних кластерів і систем автономного управління.

Отримані результати відкривають перспективи для подальших досліджень у напрямі комбінування методів підкріплювального навчання з іншими видами нейронних мереж, зокрема рекурентними та трансформерними архітектурами, для більш глибокого врахування історії змін навантаження і ресурсного стану системи. Також можливим є розширення моделі за рахунок інтеграції енергетичних обмежень, що дозволить ефективніше керувати задачами у системах із жорсткими обмеженнями на енергоспоживання, таких як мобільні та вбудовані пристрої.

Таким чином, адаптивне планування задач на основі оптимізації політик із проксимальним оновленням у поєднанні з графовими стохастичними моделями є перспективним напрямом розвитку інтелектуального управління ресурсами в складних розподілених обчислювальних середовищах, що дозволяє істотно підвищити ефективність, стійкість і масштабованість сучасних інформаційних систем.

Список використаних джерел

1. Sohani, M., & Jain, S. (2021). A predictive priority-based dynamic resource provisioning scheme with load balancing in heterogeneous cloud computing. *IEEE Access*, 9, Article 1.
2. Zhao, Z., Verma, G., Rao, C., Swami, A., & Segarra, S. (2021). Distributed scheduling using graph neural networks. In *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4720–4724). IEEE.
3. Wang, Y., Liu, X., Yang, H., Yang, X., Wang, P., Liu, Y., Luan, Z., & Qian, D. (2024). INSPIRIT: Optimizing heterogeneous task scheduling through adaptive priority in task-based runtime systems. *arXiv*. <https://arxiv.org/abs/2404.03226>