ESCALONAMENTO SENSÍVEL AO CONTEXTO NO EDGE-CLOUD CONTINUUM

Mateus Unulino dos Passos, Dhyonatan Freitas, Maurício Aronne Pillon

INTRODUÇÃO

O paradigma de *Edge-Cloud Continuum* (ECC) distribui recursos entre *Cloud*, *Fog* e *Edge*, reduzindo a latência e ampliando a resiliência de aplicações de *AI-as-a-Service* (AIaaS) voltadas à Indústria 4.0 (BELCASTRO et al., 2026). Contudo, relativamente ao posicionamento de aplicações no continuum, escalonadores padrão de orquestradores genéricos tratam *clusters* geo-distribuídos como um conjunto homogêneo de recursos (BREGAR et al., 2025), ignorando restrições e preferências de usuário como custo, latência, disponibilidade, capacidade computacional e localidade de dados. Este trabalho de Iniciação Científica teve como objetivo explorar sob escala a camada de decisão da plataforma ADAPT (FREITAS et al., 2025), um escalonador multicritério sensível ao contexto, cujo objetivo é selecionar, dentre os nós elegíveis, aquele que melhor atende às restrições e à prioridade declarada de cada carga. O resultados obtidos foram comparados ao escalonamento nativo do Kubernetes sobre o mesmo *cluster* e a mesma carga de trabalho.

DESENVOLVIMENTO

Os testes foram conduzidos em um *cluster* Kubernetes simulado com o KWOK (*Kubernetes Without Kubelet*), com 1024 nós distribuídos entre *Edge*, *Fog* e *Cloud*, com perfis de hardware (CPU, memória e GPU) derivados de dispositivos e instâncias reais. A carga de trabalho foi obtida do ‘cluster-trace-gpu-v2023’ do Grupo Alibaba (WENG et al., 2023), com 8.152 *pods*, cada um enriquecido com um perfil semântico contendo restrições (GPU, localidade dos dados) e uma prioridade de negócio (tempo real, custo reduzido, alta disponibilidade ou alto desempenho) derivada de sua classe de QoS e de seus requisitos de recursos. Um *pipeline* orientado a mensagens (MQTT) submete os *pods* ao decisor do ADAPT, que filtra os nós elegíveis por GPU, localidade e capacidade alocável disponível e, em seguida, pontua os candidatos por uma função multicritério ponderada (latência, custo, disponibilidade e capacidade computacional), com pesos distintos por prioridade e normalização vetorial (HWANG; YOON, 1981). A decisão é aplicada ao cluster via manifestos Kubernetes. Como referência (*baseline*), a mesma carga foi submetida sem preferência de camada, a cargo do escalonador padrão do Kubernetes. Cada configuração foi executada cinco vezes, com resultados agregados e visualizados em R.

RESULTADOS

A Figura 1 compara a distribuição de *pods* entre as camadas do ECC nas cinco execuções de cada estratégia. O escalonamento padrão do Kubernetes concentrou virtualmente 100% dos *pods* na *Cloud* em todas as execuções, por ausência de qualquer critério que favoreça a *Edge* ou a *Fog*. A plataforma ADAPT distribuiu os *pods* de forma heterogênea entre as camadas (aproximadamente 67% na *Fog*, 31% na *Cloud* e uma fração pequena, porém não nula, na *Edge*), refletindo os critérios de custo, latência, disponibilidade e capacidade computacional de sua função de decisão. A baixa participação da *Edge* é consistente com uma limitação estrutural do hardware modelado frente à carga utilizada. Verificou-se que há uma baixa amostragem de *pods* condizentes com cargas de trabalho para *Edge*. A Figura 2 estratifica o resultado por prioridade: o ADAPT divide aproximadamente a carga entre *Fog* e *Cloud* para ‘tempo real’, e a concentra na *Fog* para ‘custo reduzido’, enquanto o escalonamento padrão permanece insensível à prioridade, alocando sempre a totalidade dos *pods* à *Cloud*.

CONSIDERAÇÕES FINAIS

Os resultados corroboram que a camada de decisão do ADAPT diferencia efetivamente o posicionamento de cargas conforme perfis semânticos e prioridade declarada, produzindo distribuição heterogênea entre camadas do EC*C*, onde o escalonamento padrão do Kubernetes se mostrou insensível a tais critérios, concentrando-se integralmente na *Cloud*.

Palavras-chave: Edge-Cloud Continuum; Context-aware Scheduling; Kubernetes; Container Orchestration.

ANEXOS

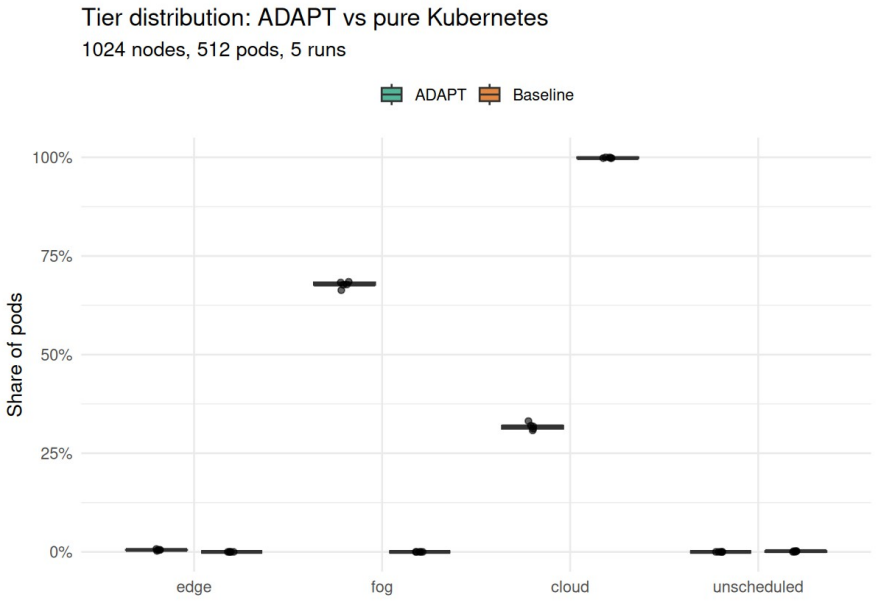


Figura 1. *Distribuição de pods entre as camadas do ECC: ADAPT versus escalonamento padrão do Kubernetes (1024 nós, 512 pods, 5 execuções).*

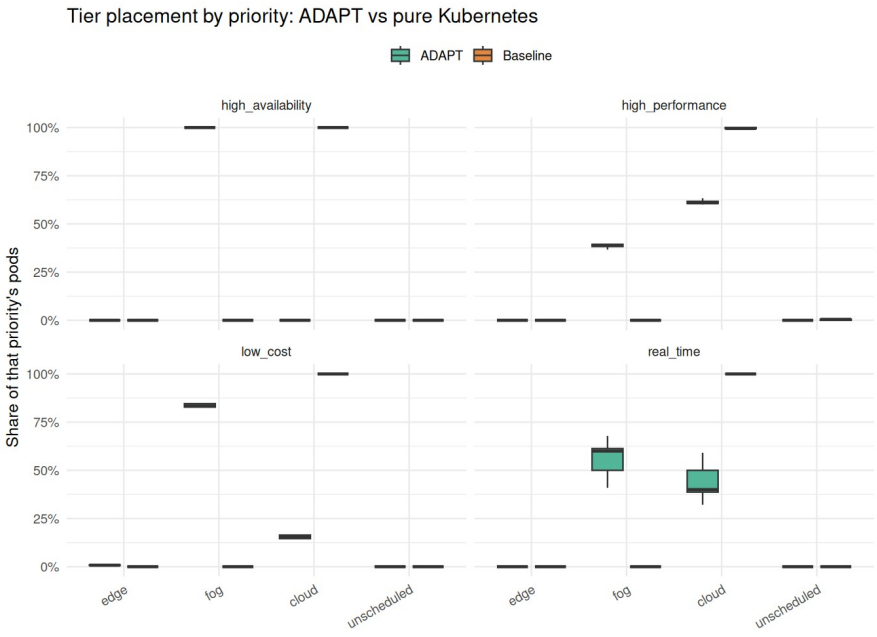


Figura 2. *Distribuição de pods por camada, estratificada por prioridade: ADAPT versus escalonamento padrão do Kubernetes.*

REFERÊNCIAS BIBLIOGRÁFICAS

BELCASTRO, Loris et al. Navigating the edge-cloud continuum: A state-of-practice survey. IEEE Access, 2026.

BREGAR, Vid; JURIC, Matjaz B. Toward Adaptive Workload Scheduling in Kubernetes Across the Edge-Cloud Continuum. 2025.

FREITAS, Dhyonatan et al. A Distributed AIaaS Platform with Context-Awareness for the Edge–Cloud Continuum Applied to Industry 4.0 (ADAPT). In: International Conference on Advanced Information Networking and Applications. Cham: Springer Nature Switzerland, 2026. p. 397-409.

KUBERNETES SIGS. KWOK: Kubernetes WithOut Kubelet. [S. l.]: Kubernetes SIGs, 2022. Disponível em: https://github.com/kubernetes-sigs/kwok. Acesso em: [DATA EM QUE VOCÊ CONSULTOU].

HWANG, C. L.; YOON, K. Multiple attribute decision making: methods and applications. Berlin: Springer-Verlag, 1981.

WENG, Q.; YANG, L.; YU, Y.; WANG, W.; TANG, X.; YANG, G.; ZHANG, L. Beware of fragmentation: scheduling GPU-sharing workloads with fragmentation gradient descent. In: USENIX