17th International Conference „Data Analysis Methods for Software Systems“
November 26 - 28, 2026, Druskininkai, Lithuania

VILNIUS TECH
Vilnius Gediminas Technical University

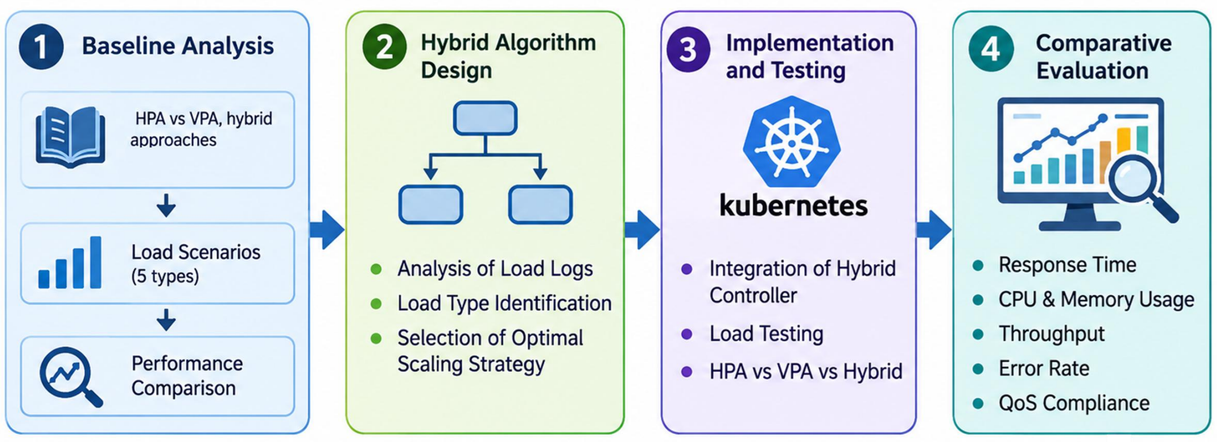
Hybrid Autoscaling Algorithm of Containerized Microservices for QoS Assurance in SaaS Systems

Dalius Mažeika, Povilas Lajus

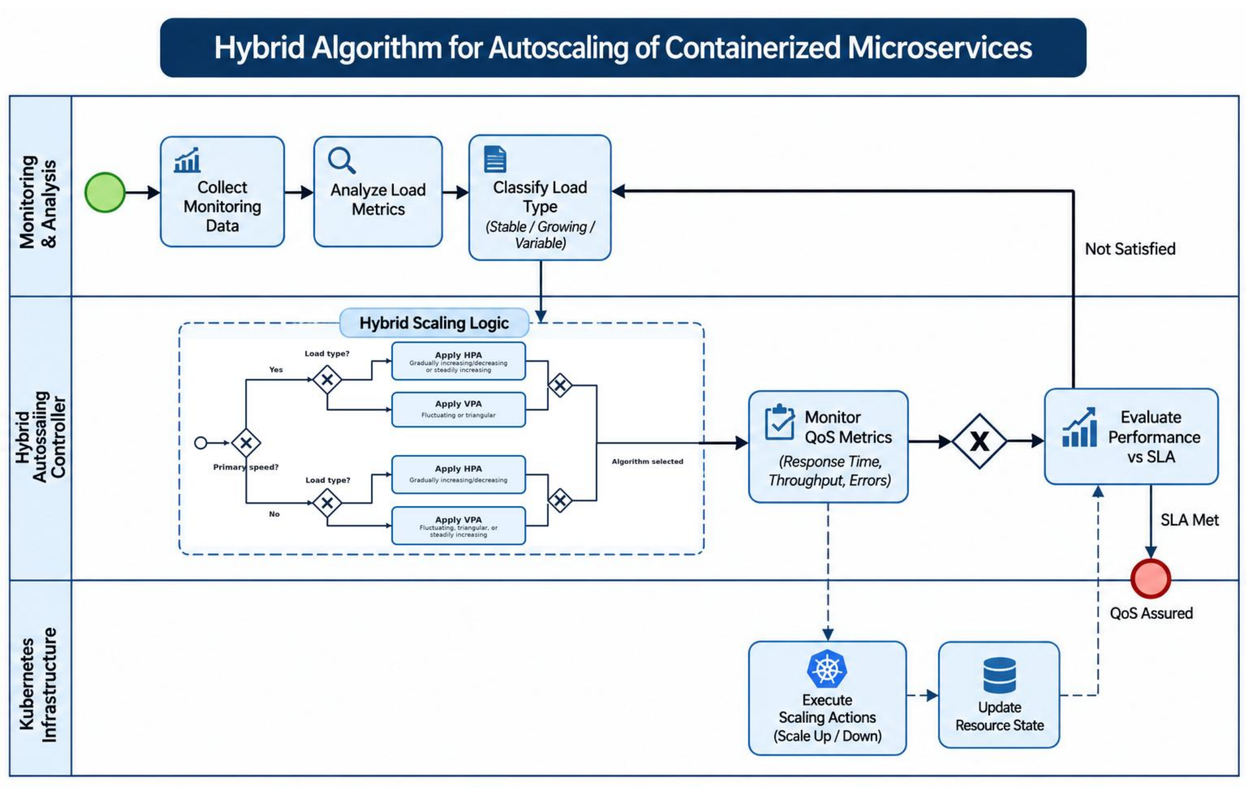
Vilnius Gediminas Technical University, Saulėtekio al. 11, Vilnius, Lithuania

Summary. QoS assurance in Kubernetes-based microservice architectures typically rely on either horizontal (HPA) or vertical (VPA) pod autoscaling, but each autoscaler has inherent limitations: HPA offers elasticity at the cost of resource overhead and cold-start delays, while VPA is resource-efficient but constrained by node capacity. A hybrid autoscaling algorithm is proposed and evaluated. It dynamically identifies incoming load patterns using a sliding-window regression model and selects the most suitable pods autoscaling method – HPA or VPA – accordingly. The model operates on real-time metrics allowing the algorithm to adapt to workload shifts without the latency. The approach was implemented and validated on a Kubernetes cluster using a CPU-intensive load under five distinct load profiles. Results show that the hybrid autoscaler reduces average response time by up to 90% and decreases error rates by up to 99.7% compared to traditional Kubernetes autoscalers, while sustaining high throughput.

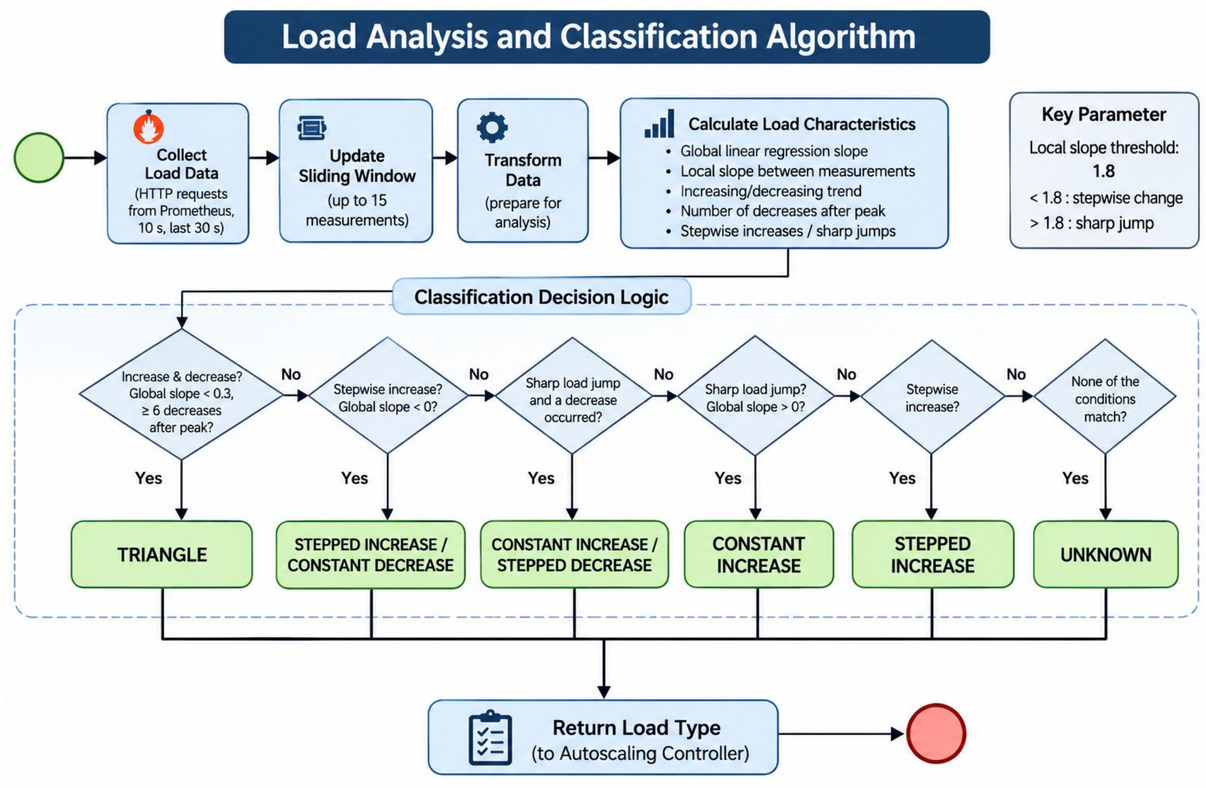
RESEARCH WORKFLOW



HYBRID ALGORITHM



LAOD CLASSIFICATION ALGORITH



RESULTS & DISCUSSION

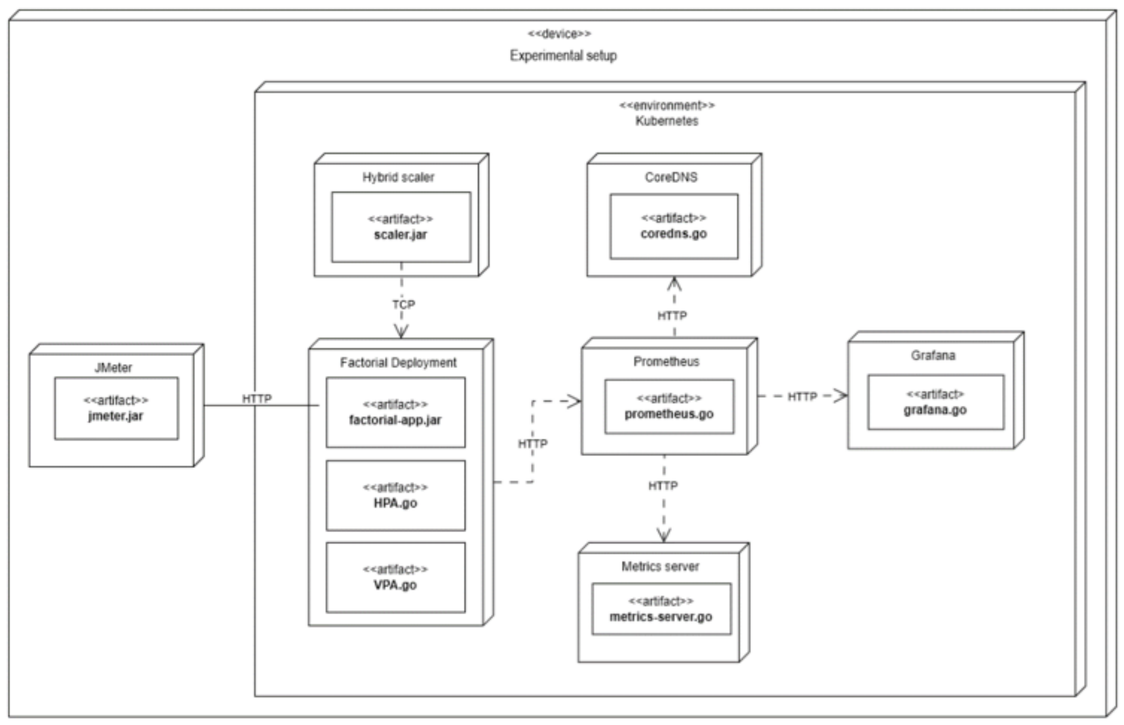


Fig 1. Experimental Setup

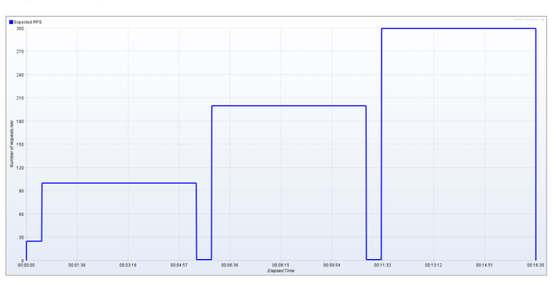


Fig 2. Increasing Load Type

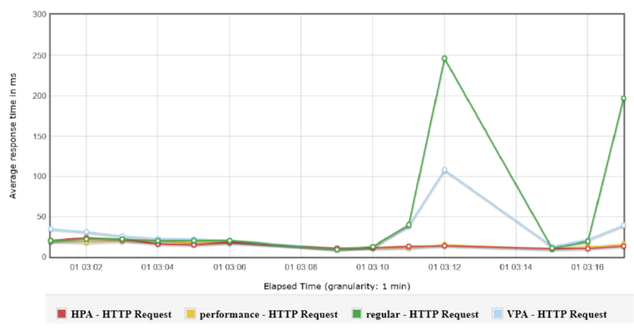


Fig 3. Response When Increasing Load is Applied

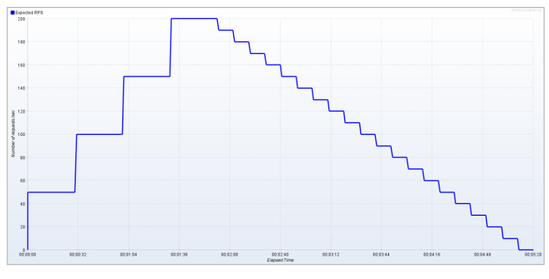


Fig 4. Increasing/Decreasing Load Type

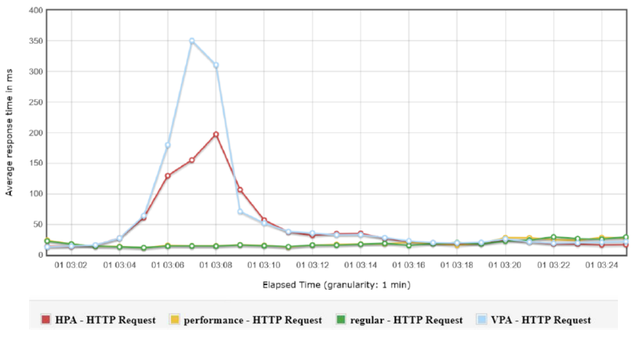


Fig 5. Increasing/Decreasing Load is Applied

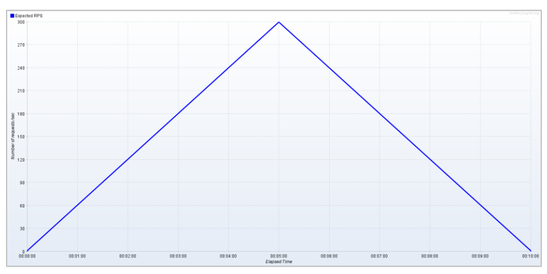


Fig 6. Triangle Load Type

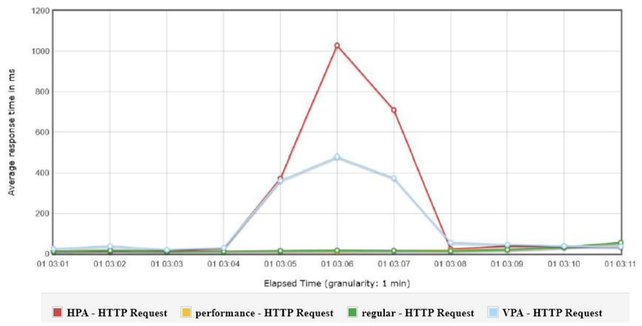


Fig 7. Response When Triangle Load is Applied

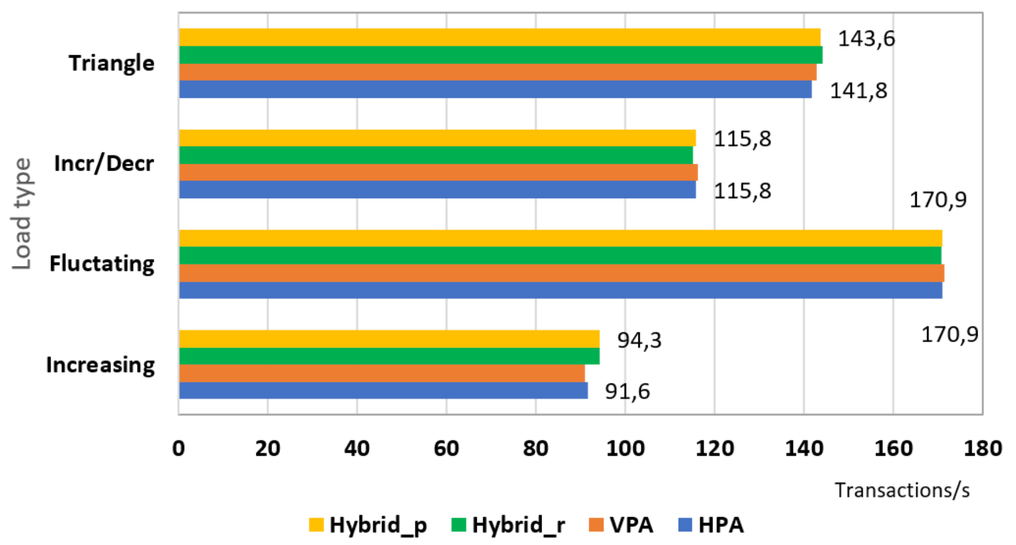


Fig 8. Throughput Comparison under Different Load