A Comparative Analysis of Traditional and Serverless Kubernetes Architectures

Xinqi Pei
School of Enterprise Computing and Digital Transformation, TU Dublin, Ireland
X00218757@myTUDublin.ie

Introduction

Cloud-native systems increasingly support real-time analytics and AI workloads that demand low latency, elastic scaling, and cost efficiency. Traditional Kubernetes Pods provide predictable performance but often suffer from idle resource waste, while serverless Kubernetes platforms (e.g. Knative) introduce event-driven scaling and pay-per-use efficiency, at the cost of cold-start latency.
This research presents an empirical comparison of traditional Kubernetes Pods and serverless Kubernetes deployments using a containerised Flask-based analytics API, instrumented with Prometheus. The study evaluates latency, CPU and memory utilisation, throughput, and cost, under steady and burst traffic patterns, to identify optimal deployment strategies for real-time analytics workloads.

Sub Topic 1

Research Questions
RQ1: How does serverless Kubernetes compare to traditional Pods in request latency?
RQ2: What are the cost-efficiency differences between both models?
RQ3: How do CPU and memory utilisation patterns differ under real-time workloads?

Sub Topic 2

Test Environment
-Kubernetes (Minikube)
-Flask-based analytics API
-Prometheus & Grafana for observability
-Docker (Python 3.10-slim)

Sub Topic 3

1.Latency Performance

- Traditional Pods deliver lower & more predictable latency
- Serverless deployments experience cold-start delays
- Latency parity observed during steady traffic

2. Resource Efficiency

- Traditional Pods show significant over-provisioning
- Serverless Pods reduce idle CPU & memory by 70–80%
- Minimal memory working set under all conditions

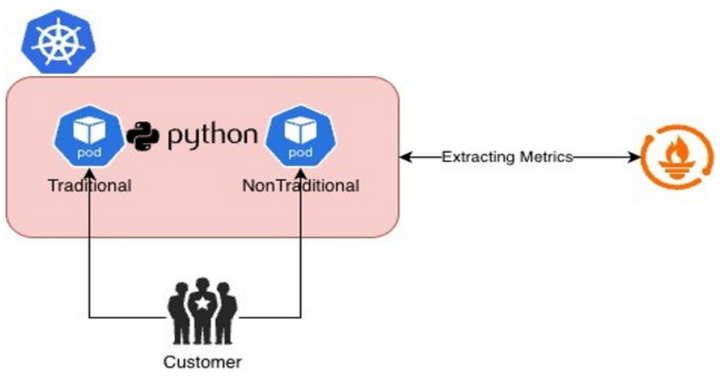
3. Cost Effectiveness

- Serverless Kubernetes yields substantially lower cost
- Best suited for bursty & intermittent workloads
- Traditional Pods remain ideal for latency-critical services

Topic Overview

This research compares traditional Kubernetes Pods and serverless Kubernetes architectures for real-time analytics workloads. Traditional Pods provide predictable performance but suffer from idle resource overhead, while serverless Kubernetes enables elastic, event-driven scaling with improved cost efficiency at the expense of potential cold-start latency. The results show that a hybrid Kubernetes–serverless approach offers the most effective balance between performance stability and resource efficiency for cloud-native analytics systems.

Figure 9: *Kubernetes Application Metrics Flow to Prometheus*



Conclusions and Future Work

The study shows that serverless Kubernetes can deliver comparable performance to traditional Pods for lightweight analytics workloads while significantly reducing resource waste and operational cost. Traditional deployments remain optimal for latency-critical services, whereas serverless models are better suited to dynamic traffic patterns. Future work will explore predictive scaling techniques, multi-node deployments, and evaluation on managed cloud platforms.

QR Code for Recording