Cloud Computing Project Report: Autoscaler Performance Comparison

Group 17 Members

1. Talal Ahmed
2. Zain Khalid
3. Usama Arif
4. Muhammad Abdullah Bin Saif

Objective

The goal of this project was to evaluate a custom autoscaler against Kubernetes Horizontal Pod Autoscaler (HPA) using two CPU utilization targets: 70% and 90%. The aim was to reduce latency and CPU usage while maintaining performance.

Experiment Setup

We ran three experiments:

- HPA 70% Target – Kubernetes scales pods when average CPU exceeds 70%.
- HPA 90% Target – Scaling happens when CPU usage goes beyond 90%.
- Custom Autoscaler – Our tailored solution that adjusts based on optimized thresholds and workload patterns.

For each setup, we tracked:

- 99th percentile latency: to measure worst-case performance.
- CPU usage: to understand how efficiently resources were used.

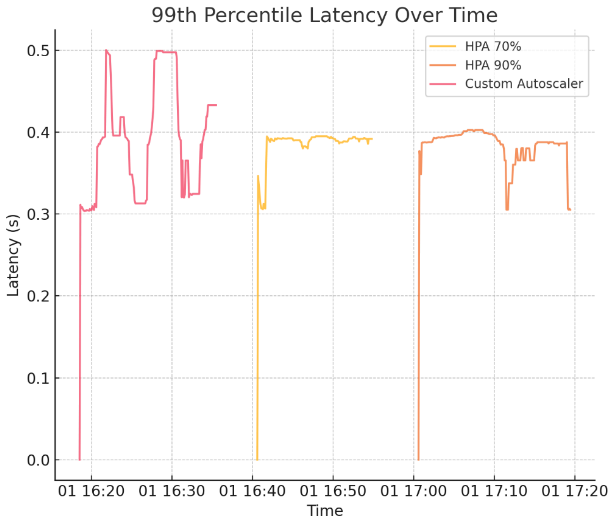
Results Summary

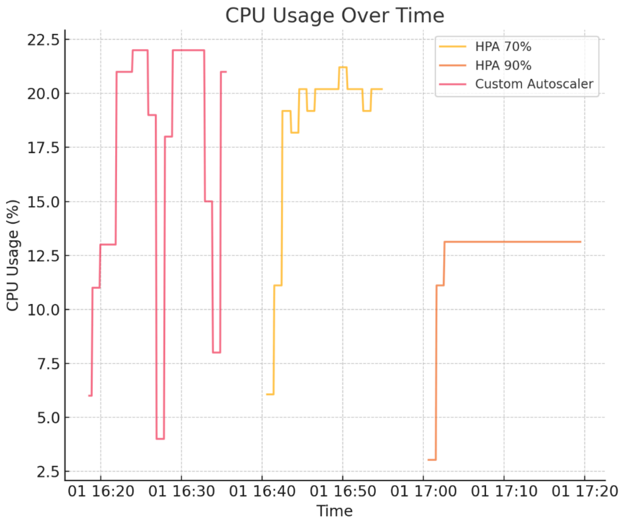
Metric	HPA 70%	HPA 90%	Custom Autoscaler
Final 99th % latency (sec)	~ 0.33	~ 0.39	~ 0.31
Final CPU usage (%)	~ 6.06%	~ 3.03%	~ 6–11% dynamic

Observations

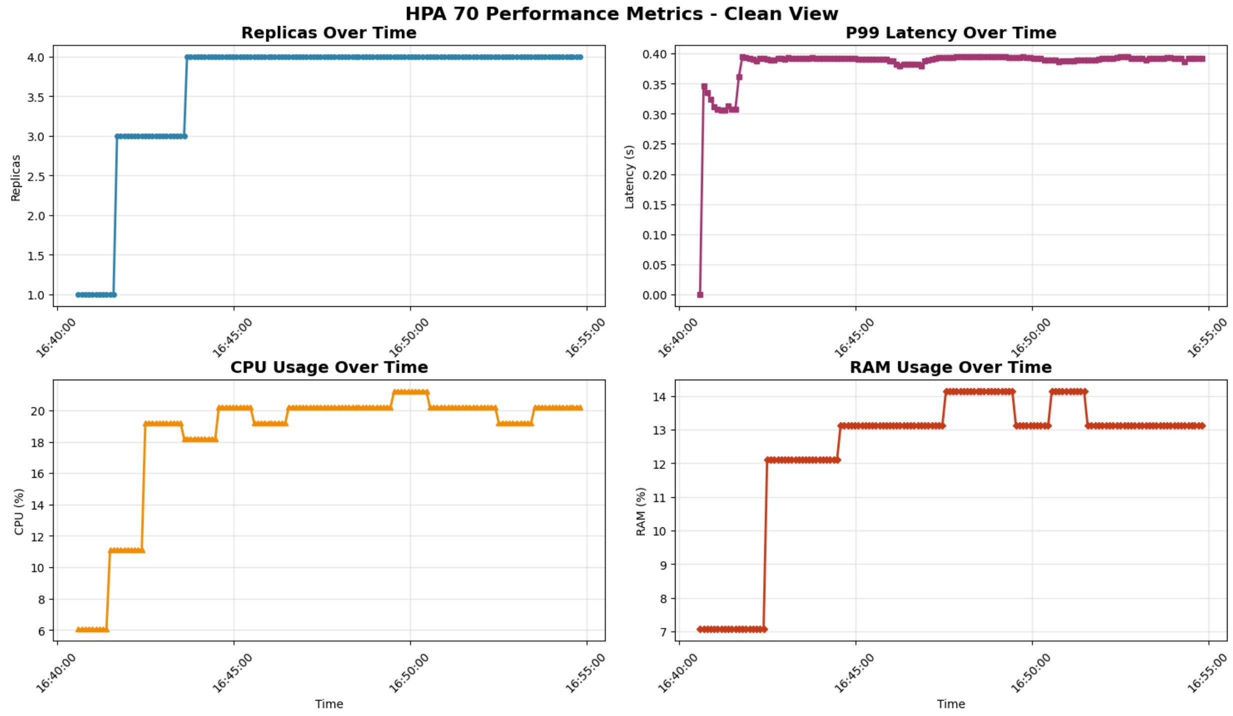
1. **Latency Comparison**
   - The custom autoscaler consistently maintained lower latency, especially under high load.
   - HPA 90% suffered from occasional latency spikes due to delayed scaling decisions.
2. **CPU Usage**
   - While HPA 90% used slightly less CPU, it came at the cost of higher latency.
   - The custom autoscaler dynamically adjusted usage, striking a better balance between performance and efficiency.
3. **Overall Performance**
   - The custom autoscaler outperformed both HPA configurations in terms of latency.
   - It demonstrated intelligent scaling with moderate CPU usage, adapting better to real-time load patterns.

Graphics

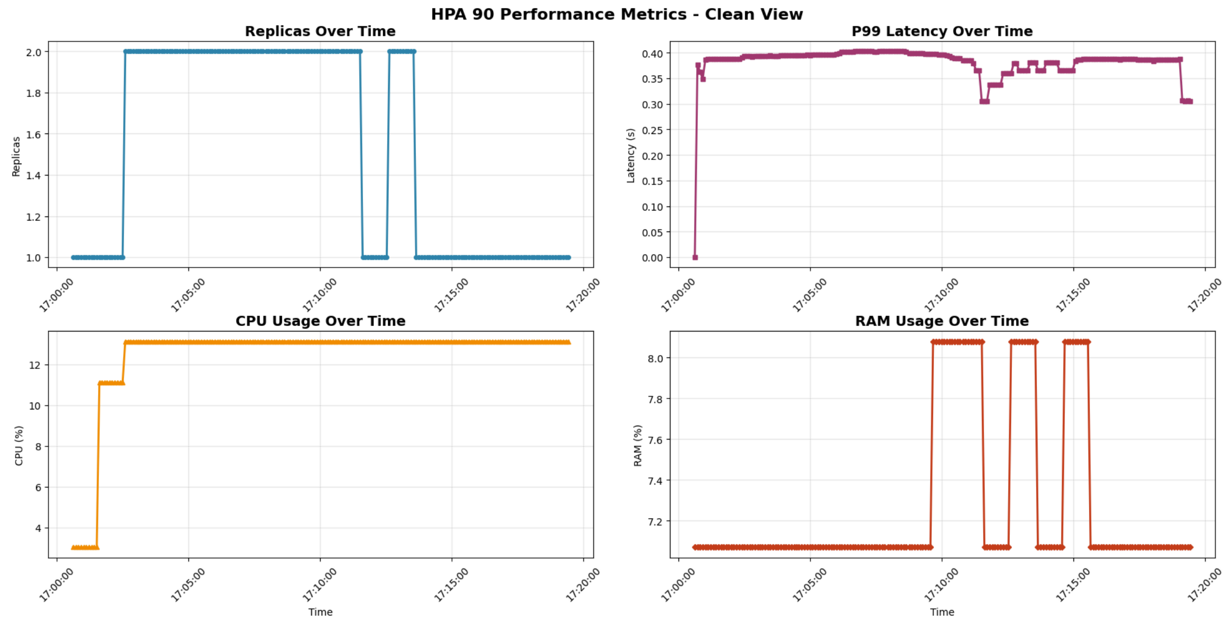




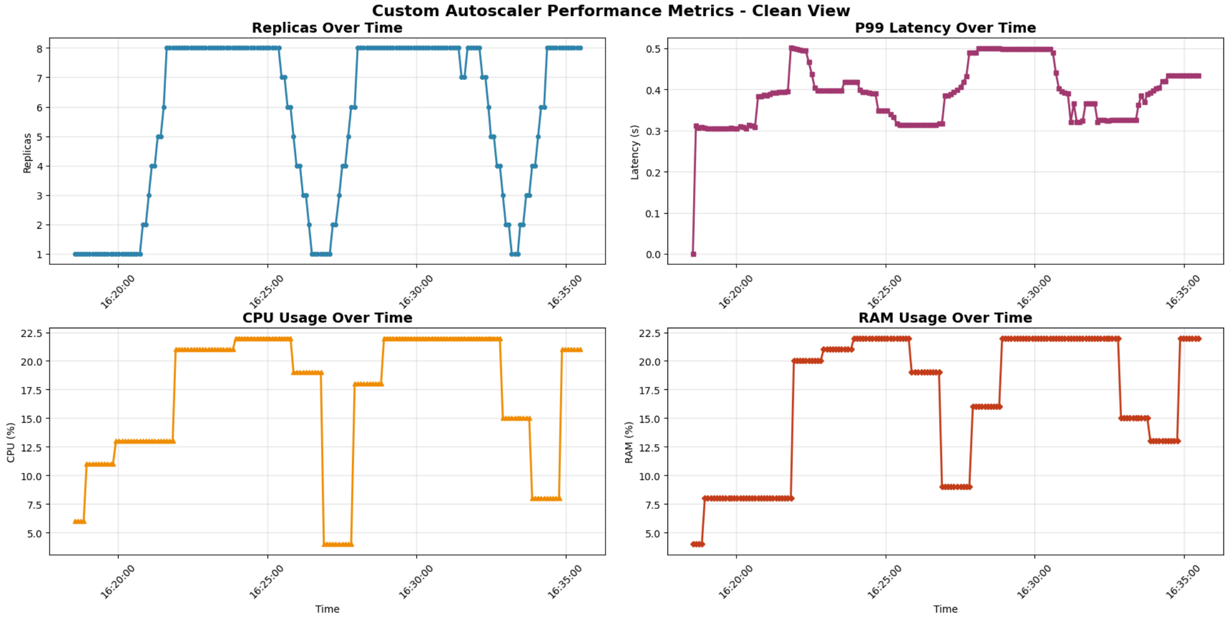
HPA 70


HPA 70 Performance Metrics - Clean View
Replicas Over Time
Replicas
1.0
1.5
2.0
2.5
3.0
3.5
4.0
16:40:00
16:45:00
16:50:00
16:55:00
P99 Latency Over Time
Latency (s)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
CPU Usage Over Time
CPU (%)
6
8
10
12
14
16
18
20
RAM Usage Over Time
RAM (%)
7
8
9
10
11
12
13
14
Time

HPA 90


HPA 90 Performance Metrics - Clean View
Replicas Over Time
Replicas
1.0
1.2
1.4
1.6
1.8
2.0
17:00:00
17:05:00
17:10:00
17:15:00
17:20:00
P99 Latency Over Time
Latency (s)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
CPU Usage Over Time
CPU (%)
4
6
8
10
12
RAM Usage Over Time
RAM (%)
7.2
7.4
7.6
7.8
8.0
Time

Custom Autoscaler


Custom Autoscaler Performance Metrics - Clean View
Replicas Over Time
Replicas
P99 Latency Over Time
Latency (s)
CPU Usage Over Time
CPU (%)
RAM Usage Over Time
RAM (%)
16:20:00
16:25:00
16:30:00
16:35:00
Time