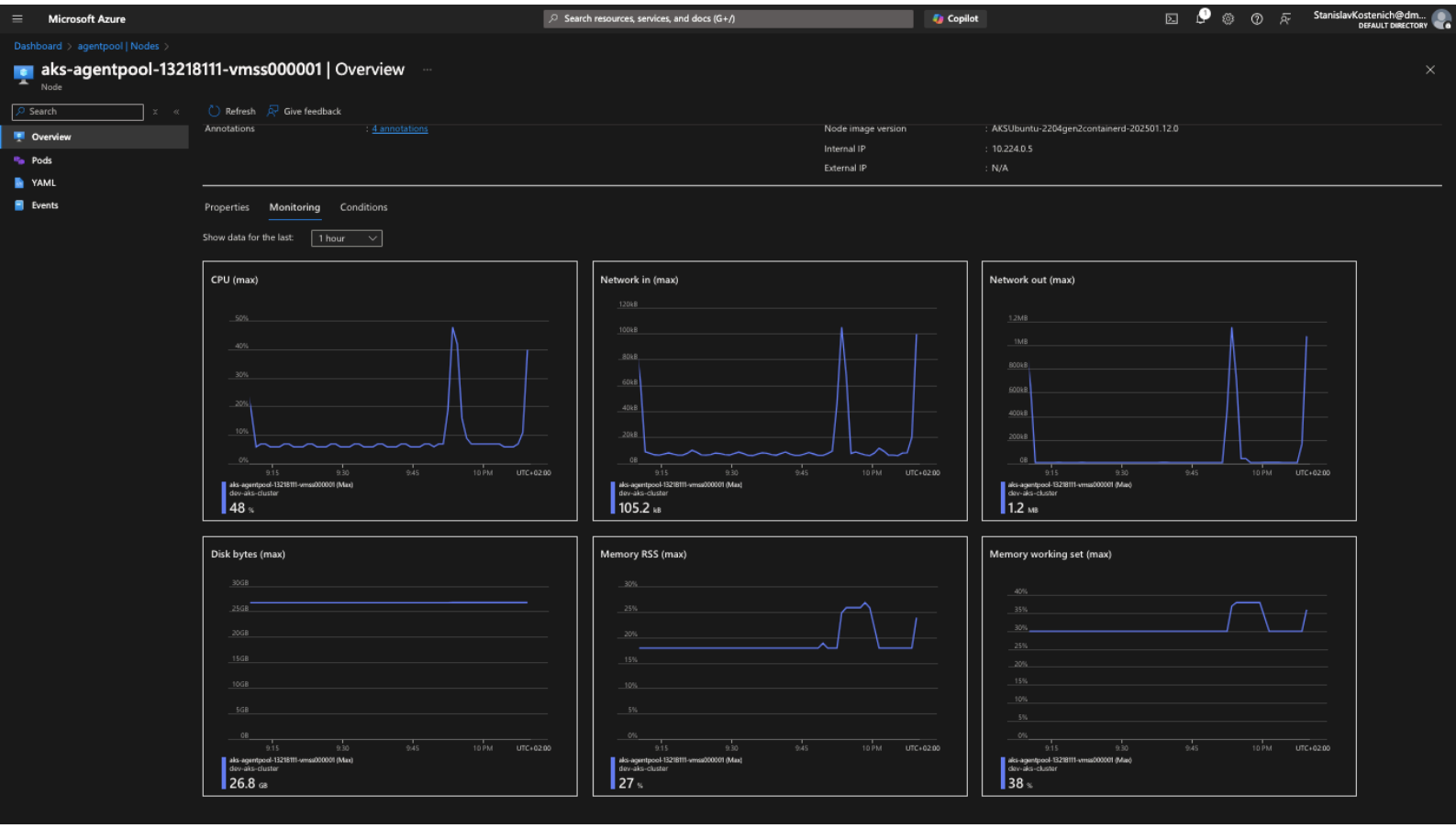
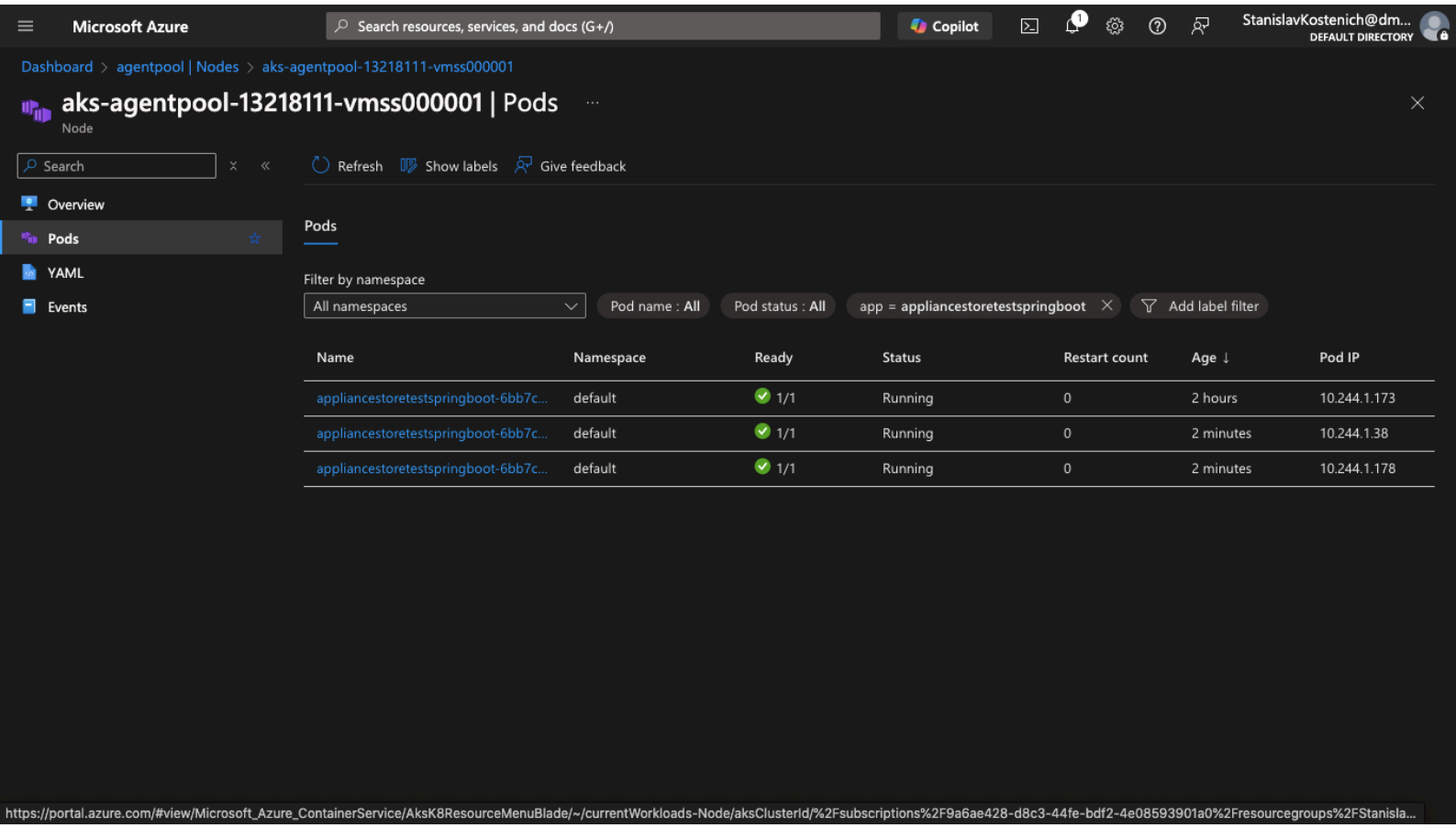
Practical Task 8: Scale Applications in AKS

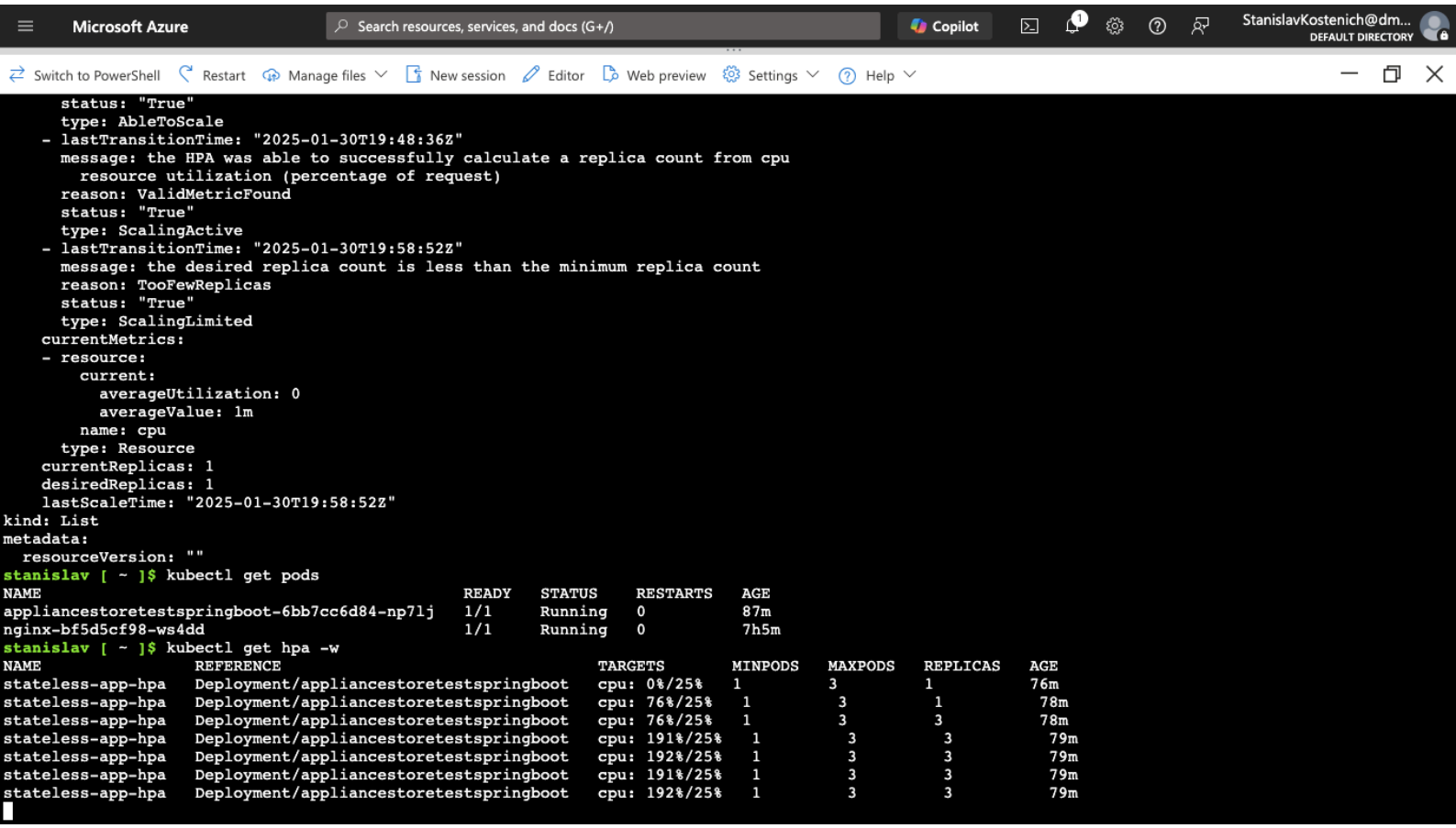
Requirements:

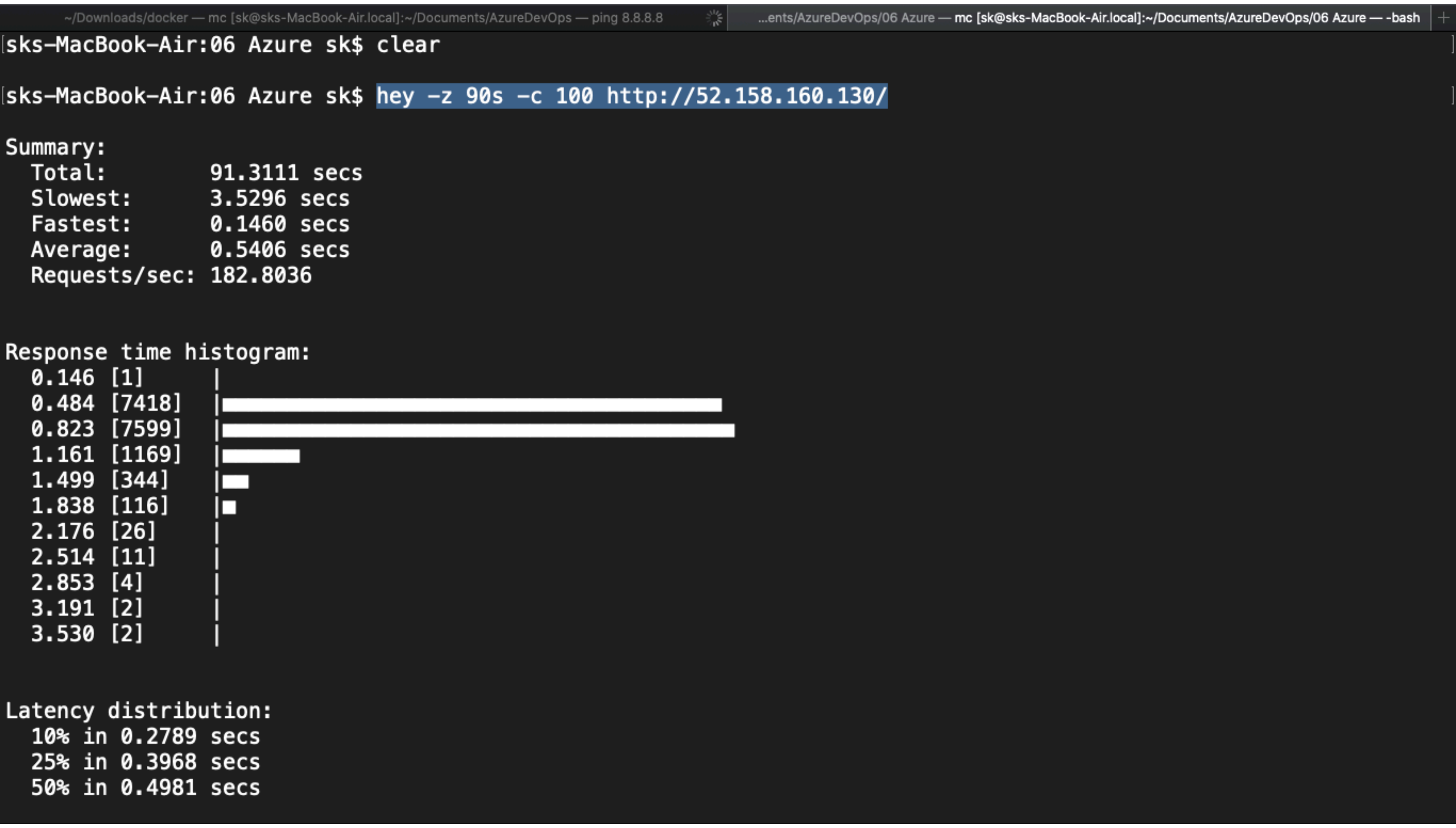
1. Deploy a stateless application to the AKS cluster using minimal resource specifications.
2. Use the kubectl scale command to manually scale the application to only 2–3 replicas for testing.
3. Set up Horizontal Pod Autoscaler (HPA) with reasonable CPU usage thresholds to minimize pod creation.

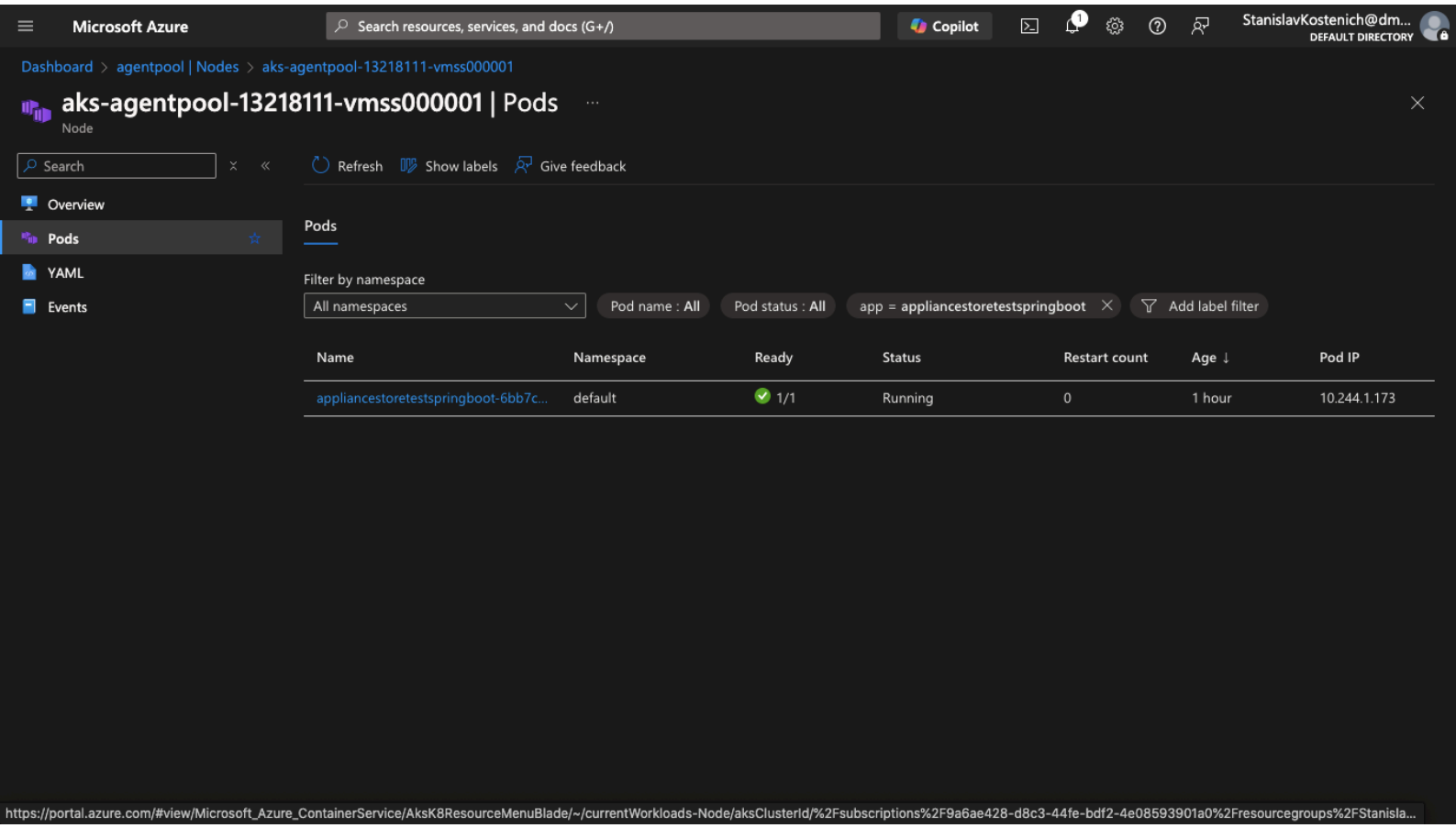
4. Simulate load on the application for a short duration and remove the deployment after observing the scaling behavior.

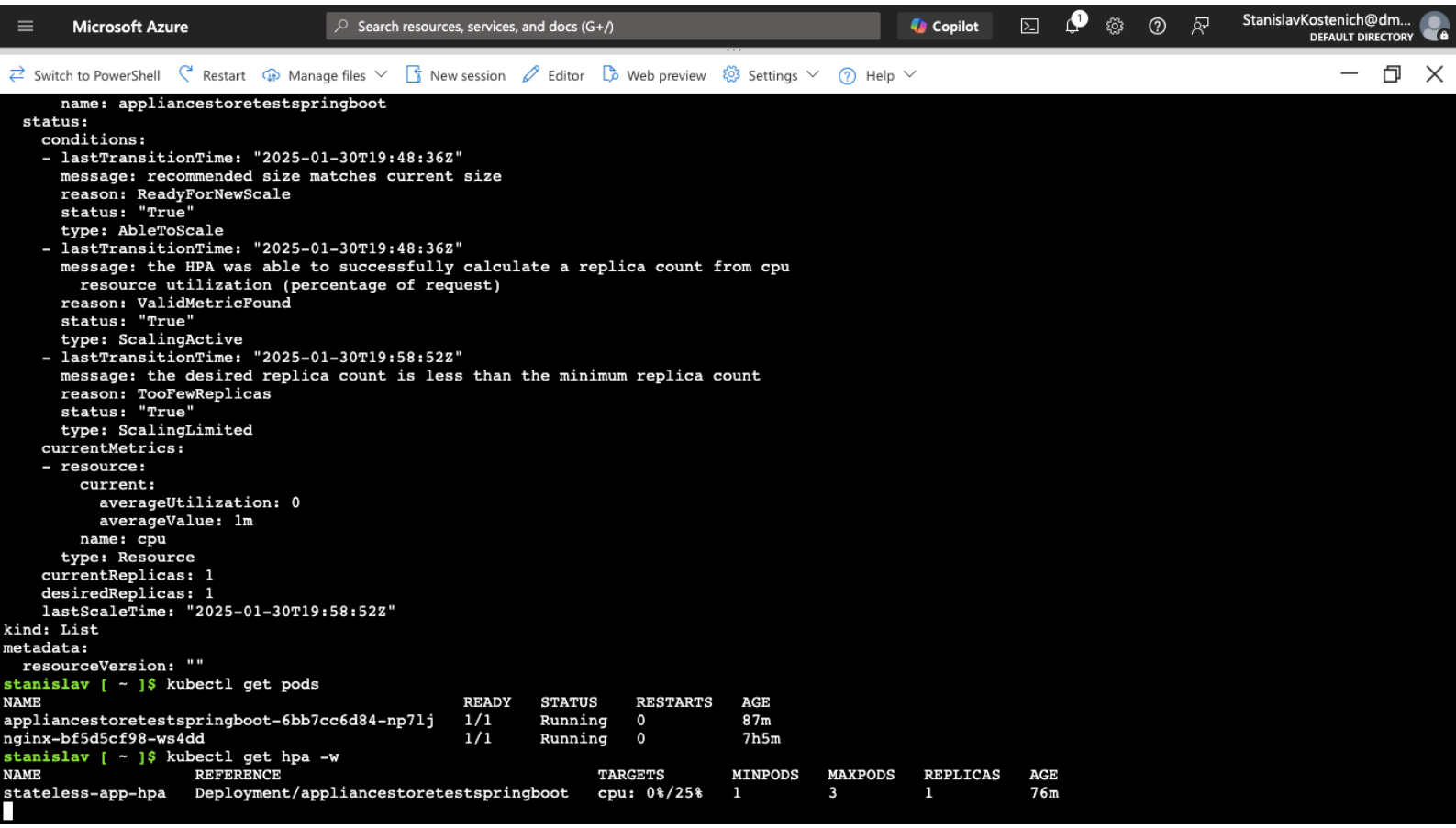

Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
StanislavKostenich@dm...
DEFAULT DIRECTORY
Dashboard > agentpool | Nodes >
aks-agentpool-13218111-vmss000001 | Overview
Node
Search
Refresh
Give feedback
Overview
Pods
YAML
Events
Annotations : 4 annotations
Node image version : AKSUbuntu-2204gen2containerd-202501.12.0
Internal IP : 10.224.0.5
External IP : N/A
Properties
Monitoring
Conditions
Show data for the last: 1 hour
CPU (max)
48 %
Network in (max)
105.2 kB
Network out (max)
1.2 MB
Disk bytes (max)
26.8 GB
Memory RSS (max)
27 %
Memory working set (max)
38 %

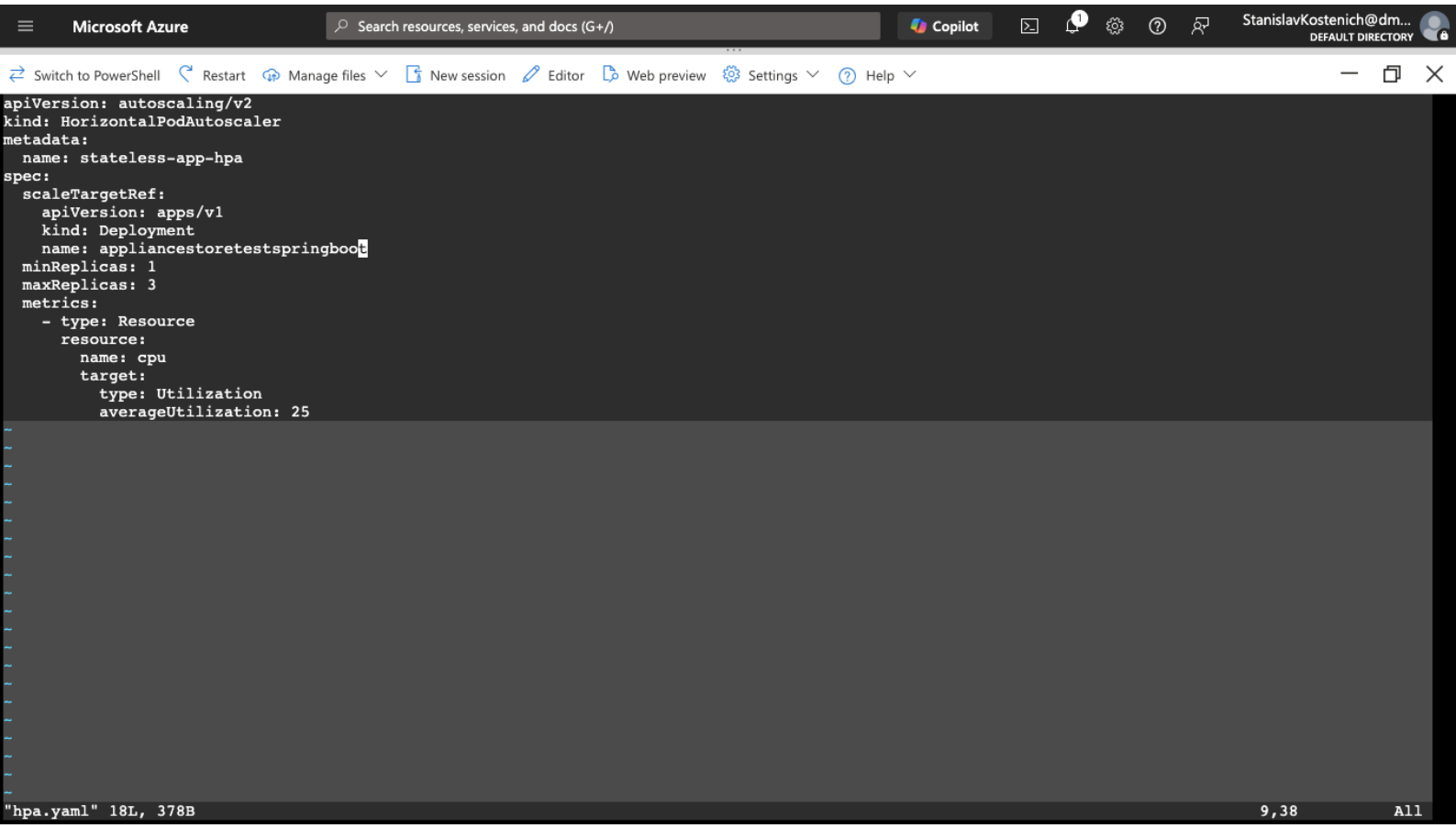



Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
StanislavKostenich@dm...
DEFAULT DIRECTORY
Switch to PowerShell
Restart
Manage files
New session
Editor
Web preview
Settings
Help
status: "True"
type: AbleToScale
- lastTransitionTime: "2025-01-30T19:48:36Z"
message: the HPA was able to successfully calculate a replica count from cpu
resource utilization (percentage of request)
reason: ValidMetricFound
status: "True"
type: ScalingActive
- lastTransitionTime: "2025-01-30T19:58:52Z"
message: the desired replica count is less than the minimum replica count
reason: TooFewReplicas
status: "True"
type: ScalingLimited
currentMetrics:
- resource:
current:
averageUtilization: 0
averageValue: 1m
name: cpu
type: Resource
currentReplicas: 1
desiredReplicas: 1
lastScaleTime: "2025-01-30T19:58:52Z"
kind: List
metadata:
resourceVersion: ""
stanislav [~]$ kubectl get pods
NAME READY STATUS RESTARTS AGE
appliancestoretestspringboot-6bb7cc6d84-np71j 1/1 Running 0 87m
nginx-bf5d5cf98-ws4dd 1/1 Running 0 7h5m
stanislav [~]$ kubectl get hpa -w
NAME REFERENCE TARGETS MINPODS MAXPODS REPLICAS AGE
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 0%/25% 1 3 1 76m
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 76%/25% 1 3 1 78m
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 76%/25% 1 3 3 78m
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 191%/25% 1 3 3 79m
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 192%/25% 1 3 3 79m
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 191%/25% 1 3 3 79m
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 192%/25% 1 3 3 79m


~/Downloads/docker — mc [sk@sks-MacBook-Air.local]:~/Documents/AzureDevOps — ping 8.8.8.8
...ents/AzureDevOps/06 Azure — mc [sk@sks-MacBook-Air.local]:~/Documents/AzureDevOps/06 Azure — -bash
sks-MacBook-Air:06 Azure sk$ clear
sks-MacBook-Air:06 Azure sk$ hey -z 90s -c 100 http://52.158.160.130/
Summary:
Total: 91.3111 secs
Slowest: 3.5296 secs
Fastest: 0.1460 secs
Average: 0.5406 secs
Requests/sec: 182.8036
Response time histogram:
0.146 [1]
0.484 [7418]
0.823 [7599]
1.161 [1169]
1.499 [344]
1.838 [116]
2.176 [26]
2.514 [11]
2.853 [4]
3.191 [2]
3.530 [2]
Latency distribution:
10% in 0.2789 secs
25% in 0.3968 secs
50% in 0.4981 secs


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
StanislavKostenich@dm...
DEFAULT DIRECTORY
Dashboard > agentpool | Nodes > aks-agentpool-13218111-vmss000001
aks-agentpool-13218111-vmss000001 | Pods
Node
Search
Refresh
Show labels
Give feedback
Overview
Pods
YAML
Events
Pods
Filter by namespace
All namespaces
Pod name : All
Pod status : All
app = appliancestoretestspringboot
Add label filter
Name
Namespace
Ready
Status
Restart count
Age
Pod IP
appliancestoretestspringboot-6bb7c...
default
1/1
Running
0
1 hour
10.244.1.173
https://portal.azure.com/#view/Microsoft_Azure_ContainerService/AksK8ResourceMenuBlade/~/currentWorkloads-Node/aksClusterId/%2Fsubscriptions%2F9a6ae428-d8c3-44fe-bdf2-4e08593901a0%2Fresourcegroups%2FStanisla...


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
StanislavKostenich@dm...
DEFAULT DIRECTORY
Switch to PowerShell
Restart
Manage files
New session
Editor
Web preview
Settings
Help
name: appliancestoretestspringboot
status:
conditions:
- lastTransitionTime: "2025-01-30T19:48:36Z"
message: recommended size matches current size
reason: ReadyForNewScale
status: "True"
type: AbleToScale
- lastTransitionTime: "2025-01-30T19:48:36Z"
message: the HPA was able to successfully calculate a replica count from cpu
resource utilization (percentage of request)
reason: ValidMetricFound
status: "True"
type: ScalingActive
- lastTransitionTime: "2025-01-30T19:58:52Z"
message: the desired replica count is less than the minimum replica count
reason: TooFewReplicas
status: "True"
type: ScalingLimited
currentMetrics:
- resource:
current:
averageUtilization: 0
averageValue: 1m
name: cpu
type: Resource
currentReplicas: 1
desiredReplicas: 1
lastScaleTime: "2025-01-30T19:58:52Z"
kind: List
metadata:
resourceVersion: ""
stanislav [~]$ kubectl get pods
NAME READY STATUS RESTARTS AGE
appliancestoretestspringboot-6bb7cc6d84-np7lj 1/1 Running 0 87m
nginx-bf5d5cf98-ws4dd 1/1 Running 0 7h5m
stanislav [~]$ kubectl get hpa -w
NAME REFERENCE TARGETS MINPODS MAXPODS REPLICAS AGE
stateless-app-hpa Deployment/appliancestoretestspringboot cpu: 0%/25% 1 3 1 76m


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
StanislavKostenich@dm...
DEFAULT DIRECTORY
Switch to PowerShell
Restart
Manage files
New session
Editor
Web preview
Settings
Help
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
name: stateless-app-hpa
spec:
scaleTargetRef:
apiVersion: apps/v1
kind: Deployment
name: appliancestoretestspringboot
minReplicas: 1
maxReplicas: 3
metrics:
- type: Resource
resource:
name: cpu
target:
type: Utilization
averageUtilization: 25
"hpa.yaml" 18L, 378B
9,38
All

4. Remove the ConfigMap, Secret, and deployment after testing.

Practical Task 8: Scale Applications in AKS

Requirements:

1. Deploy a stateless application to the AKS cluster using minimal resource specifications.
2. Use the kubectl scale command to manually scale the application to only 2–3 replicas for testing.
3. Set up Horizontal Pod Autoscaler (HPA) with reasonable CPU usage thresholds to minimize pod creation.

4. Simulate load on the application for a short duration and remove the deployment after observing the scaling behavior.