Microsoft Azure

쿠버네티스, DevOps에서부터 Production 환경까지

at Open Infrastructure Community Day Korea 2020

최영락, 박인혜 / 마이크로소프트

최영락

Developer Product Marketing Manager at Microsoft

@ianychoi

About Me…

Client 개발

Server 개발

오픈소스 개발

Cloud 관리 플랫폼 개발

Container 관리 플랫폼 개발

Cloud Infra Architecture

Application & DevOps Architecture

mail: cosmos0703@gmail.com

Microsoft

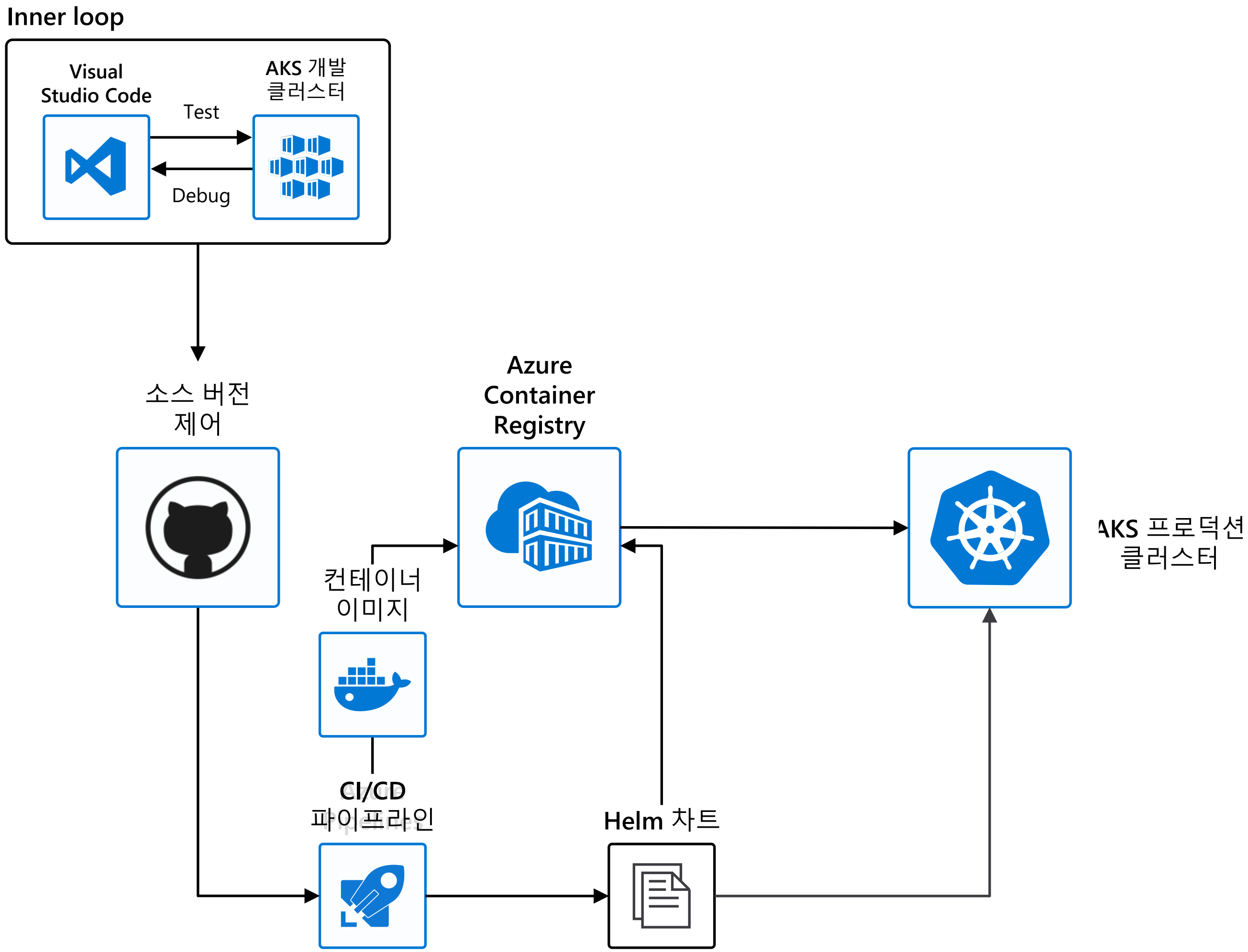
Azure

1. 쿠버네티스와 DevOps

DevOps와 함께 컨테이너 개발 속도 내기

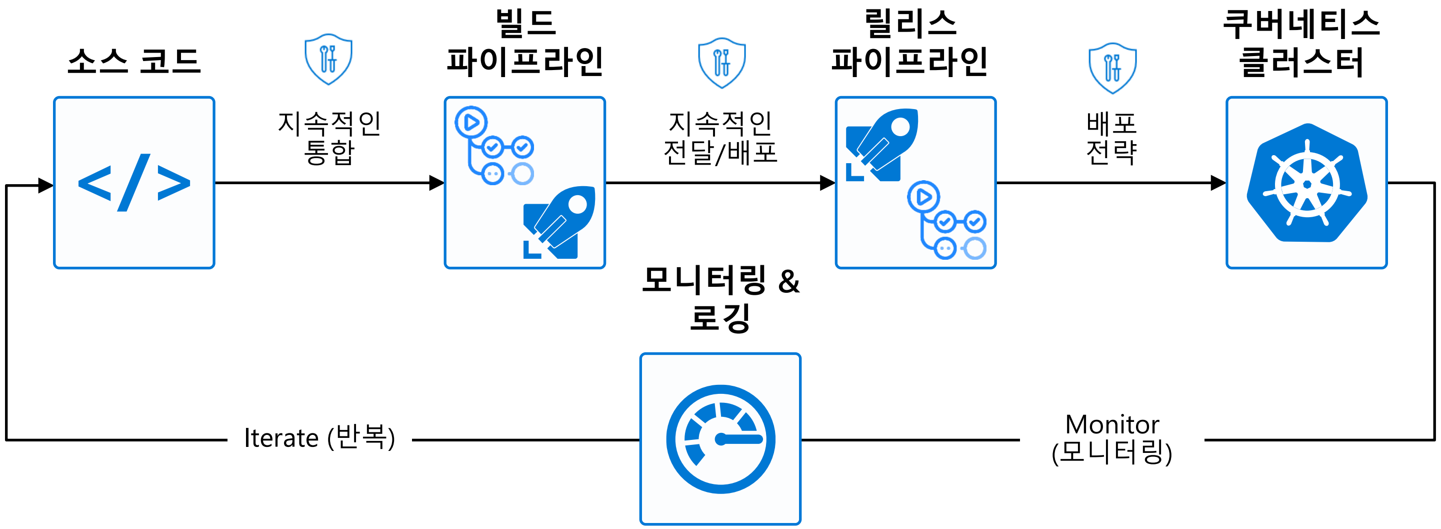
CI/CD 통합을 통해 팀에서 반복되는 Kubernetes 개발 경험에 도움되는 Azure 클라우드 관련 사항

- Visual Studio Code: 네이티브 컨테이너 및 Kubernetes 개발 환경 지원
- Helm을 지원하는 프라이빗 컨테이너 레지스트리
- 의존성을 별도로 분리하지 않고 Kubernetes 앱 개발 및 테스트
- 몇 번의 클릭만으로 자동화된 작업을 CI/CD 파이브라인을 통해 코드 머지, 컨테이너화를 효율적으로 수행
- 미리 구성된 카나리 배포 전략 수행
- 심도 있는 빌드 및 배포 프로세스를 리뷰 및 통합테스트와 연계하여 가능

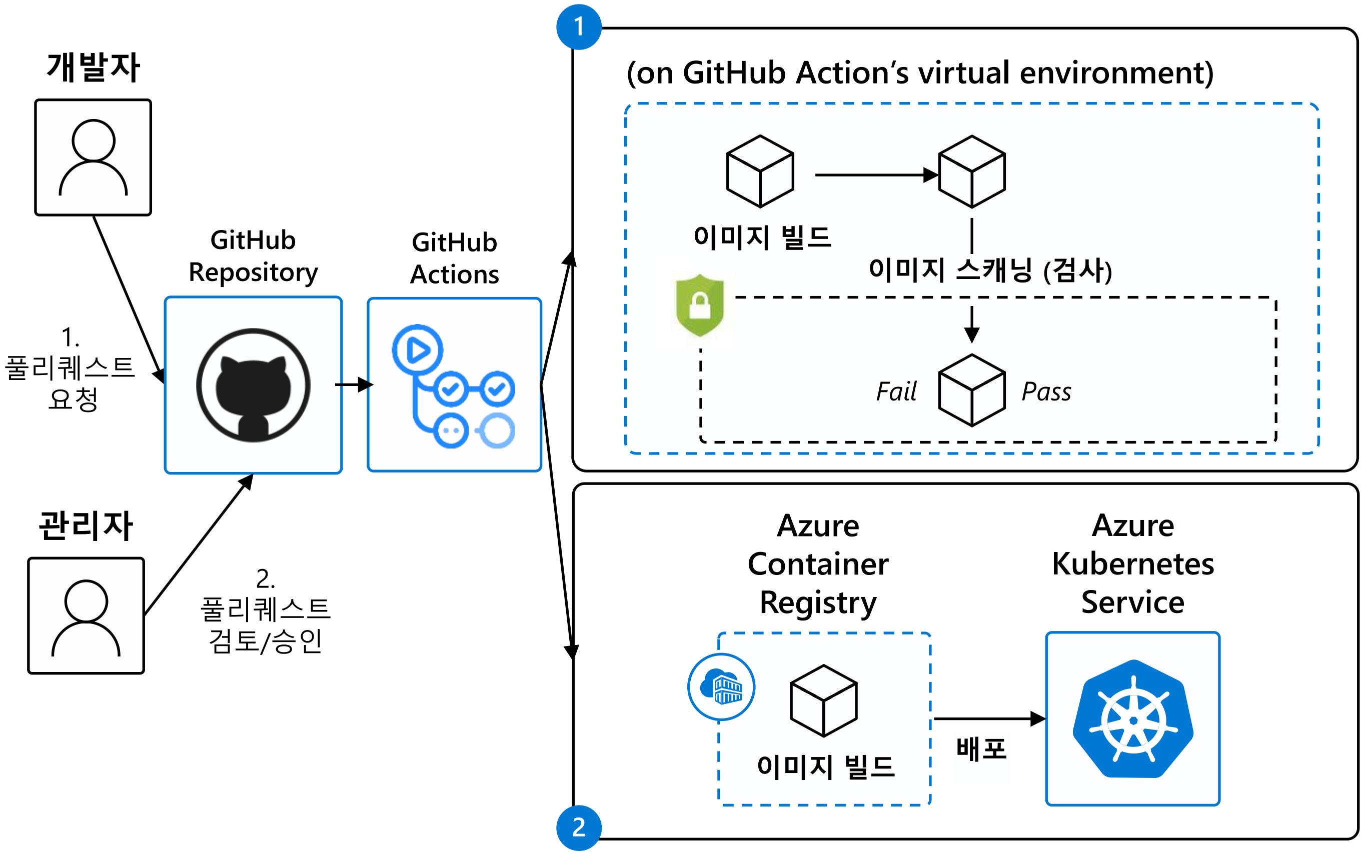


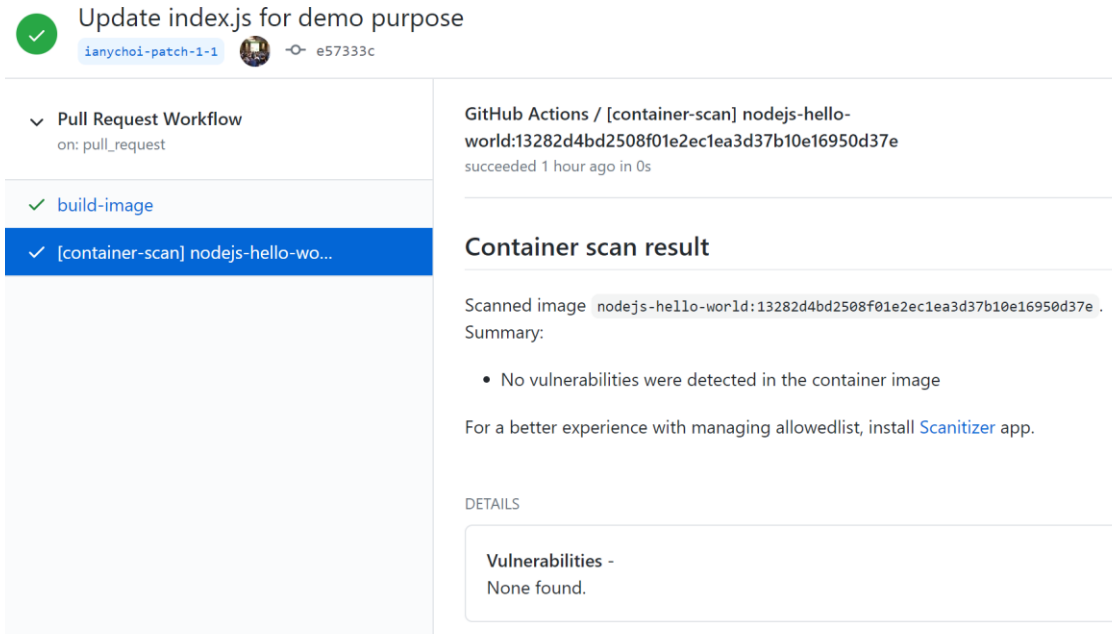
안전한 (Secure) DevOps

- 쿠버네티스 및 CI/CD와 함께 코드를 보다 빠르게 배포/전달
- 지속적인 모니터링을 통한 피드백 수용 과정을 가속화
- 지속적인 보안 & 심층 추적을 기반으로 한 **속도 (speed)**와 **보안**에 대한 **밸런싱**



데모 환경: GitHub Actions for Kubernetes on Azure





Update index.js for demo purpose

ianychoi-patch-1-1 e57333c

Pull Request Workflow
on: pull_request

✓ build-image

✓ [container-scan] nodejs-hello-wo...

GitHub Actions / [container-scan] nodejs-hello-world:13282d4bd2508f01e2ec1ea3d37b10e16950d37e
succeeded 1 hour ago in 0s

Container scan result

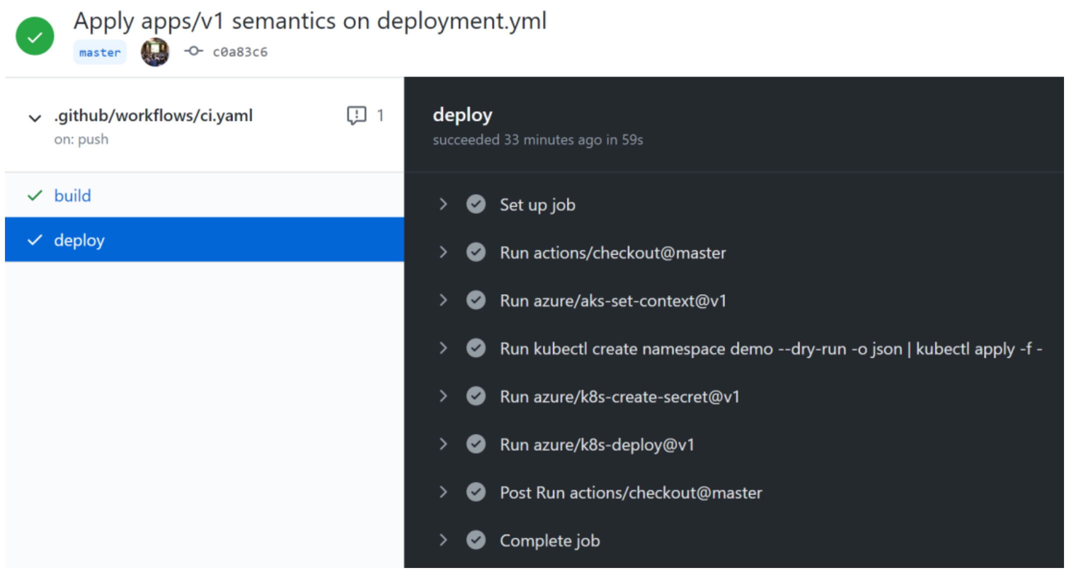
Scanned image `nodejs-hello-world:13282d4bd2508f01e2ec1ea3d37b10e16950d37e`.
Summary:

- No vulnerabilities were detected in the container image

For a better experience with managing allowedlist, install Scanitizer app.

DETAILS

Vulnerabilities -
None found.



Apply apps/v1 semantics on deployment.yml

master c0a83c6

.github/workflows/ci.yaml 1
on: push

✓ build

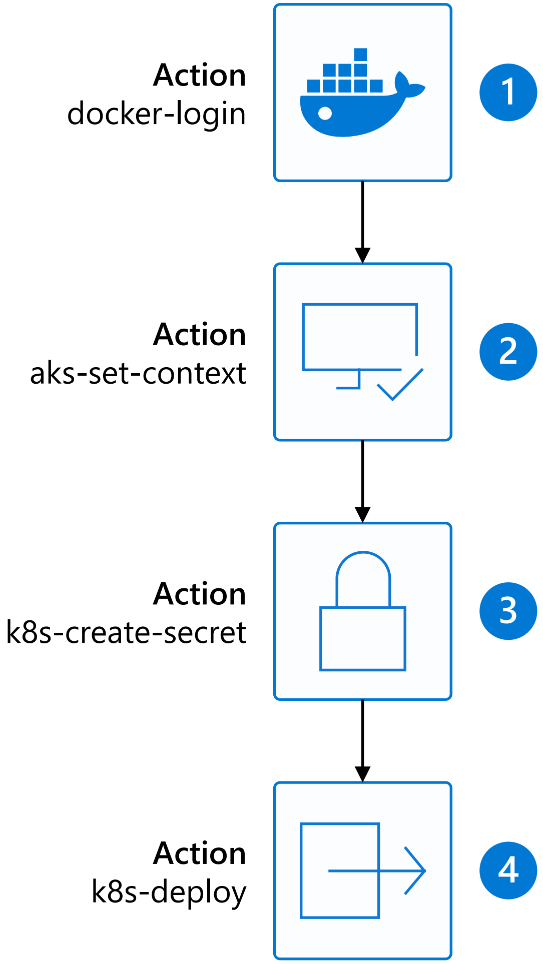
✓ deploy

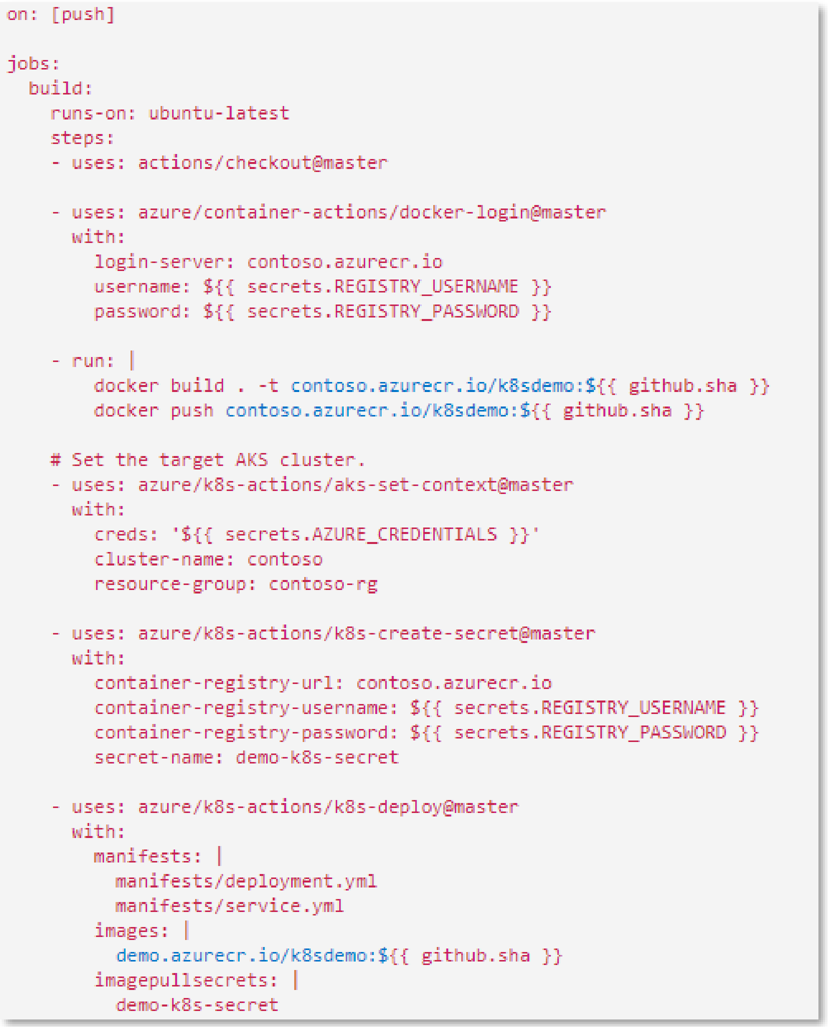
deploy
succeeded 33 minutes ago in 59s

- Set up job
- Run actions/checkout@master
- Run azure/aks-set-context@v1
- Run kubectl create namespace demo --dry-run -o json | kubectl apply -f -
- Run azure/k8s-create-secret@v1
- Run azure/k8s-deploy@v1
- Post Run actions/checkout@master
- Complete job

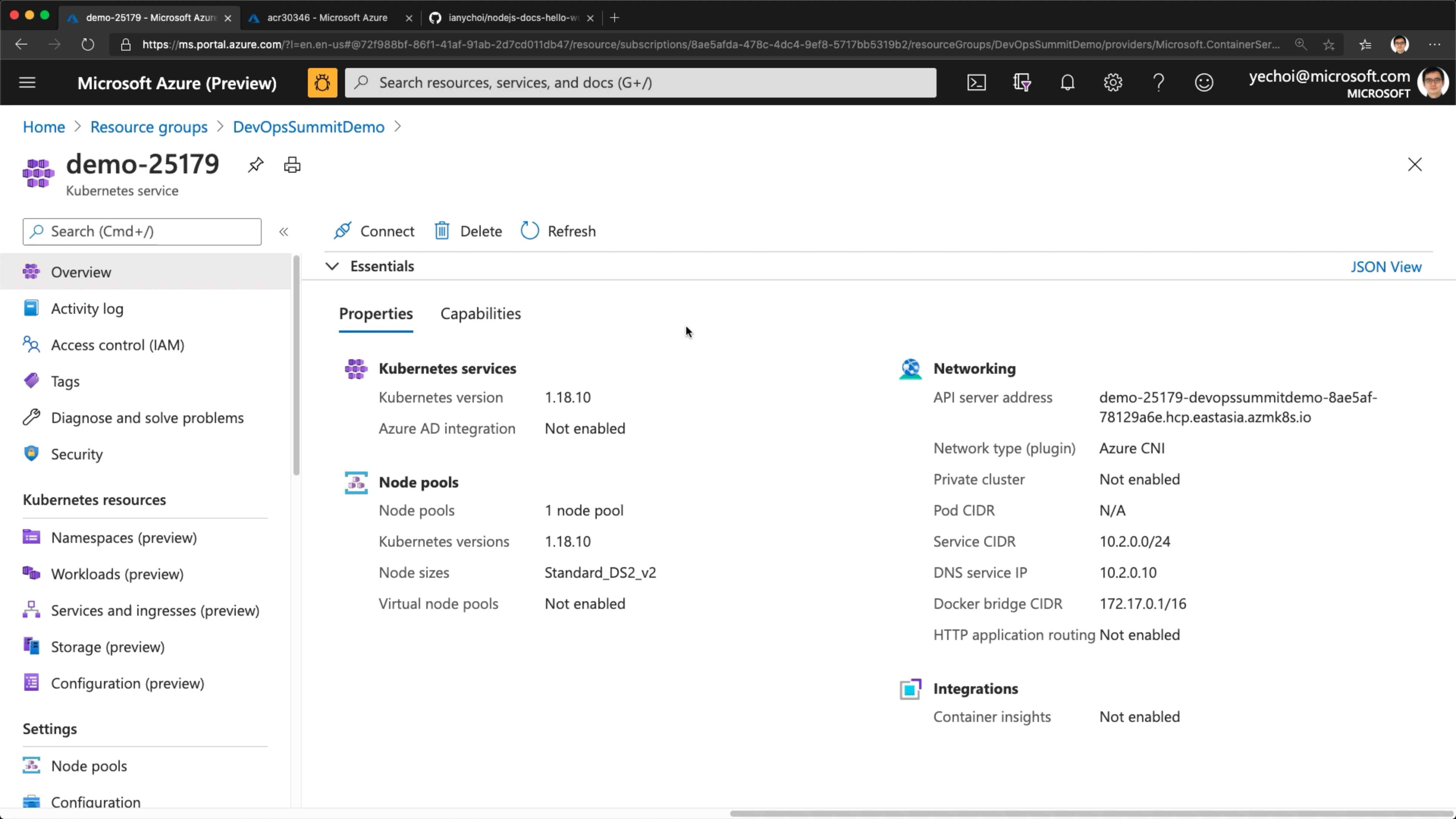
데모 설명: GitHub Actions for Kubernetes on Azure

1. 애저 구독에 안전하게 인증 & 로그인 수행
2. 대상 AKS 클러스터 지정
3. 민감한 정보를 안전하게 관리하기 위해 쿠버네티스 시크릿 (Secret) 오브젝트 생성
4. 쿠버네티스 클러스터에 연결 후 manifest 등을 배포





```yaml
on: [push]

jobs:
  build:
    runs-on: ubuntu-latest
    steps:
    - uses: actions/checkout@master

    - uses: azure/container-actions/docker-login@master
      with:
        login-server: contoso.azurecr.io
        username: ${{ secrets.REGISTRY_USERNAME }}
        password: ${{ secrets.REGISTRY_PASSWORD }}

    - run: |
        docker build . -t contoso.azurecr.io/k8sdemo:${{ github.sha }}
        docker push contoso.azurecr.io/k8sdemo:${{ github.sha }}

    # Set the target AKS cluster.
    - uses: azure/k8s-actions/aks-set-context@master
      with:
        creds: '${{ secrets.AZURE_CREDENTIALS }}'
        cluster-name: contoso
        resource-group: contoso-rg

    - uses: azure/k8s-actions/k8s-create-secret@master
      with:
        container-registry-url: contoso.azurecr.io
        container-registry-username: ${{ secrets.REGISTRY_USERNAME }}
        container-registry-password: ${{ secrets.REGISTRY_PASSWORD }}
        secret-name: demo-k8s-secret

    - uses: azure/k8s-actions/k8s-deploy@master
      with:
        manifests: |
          manifests/deployment.yml
          manifests/service.yml
        images: |
          demo.azurecr.io/k8sdemo:${{ github.sha }}
        imagepullsecrets: |
          demo-k8s-secret
```


demo-25179 - Microsoft Azure
acr30346 - Microsoft Azure
ianychoi/nodejs-docs-hello-w
https://ms.portal.azure.com/?l=en.en-us#@72f988bf-86f1-41af-91ab-2d7cd011db47/resource/subscriptions/8ae5afda-478c-4dc4-9ef8-5717bb5319b2/resourceGroups/DevOpsSummitDemo/providers/Microsoft.ContainerSer...
Microsoft Azure (Preview)
Search resources, services, and docs (G+/)
yechoi@microsoft.com
MICROSOFT
Home > Resource groups > DevOpsSummitDemo >
Search (Cmd+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Security
Kubernetes resources
Namespaces (preview)
Workloads (preview)
Services and ingresses (preview)
Storage (preview)
Configuration (preview)
Settings
Node pools
Configuration
Connect
Delete
Refresh
Essentials
JSON View
Properties
Capabilities

demo-25179

Kubernetes service

Kubernetes services

Kubernetes version	1.18.10
Azure AD integration	Not enabled

Node pools

Node pools	1 node pool
Kubernetes versions	1.18.10
Node sizes	Standard_DS2_v2
Virtual node pools	Not enabled

Networking

API server address	demo-25179-devopssummitdemo-8ae5af-78129a6e.hcp.eastasia.azmk8s.io
Network type (plugin)	Azure CNI
Private cluster	Not enabled
Pod CIDR	N/A
Service CIDR	10.2.0.0/24
DNS service IP	10.2.0.10
Docker bridge CIDR	172.17.0.1/16
HTTP application routing	Not enabled

Integrations

Container insights	Not enabled

Microsoft & 커뮤니티 뉴스

정기적인 커뮤니티 & 마이크로소프트 소식 및 커뮤니티 이벤트, 워크샵에 대한 안내를 받으실 수 있습니다.

https://aka.ms/devKR

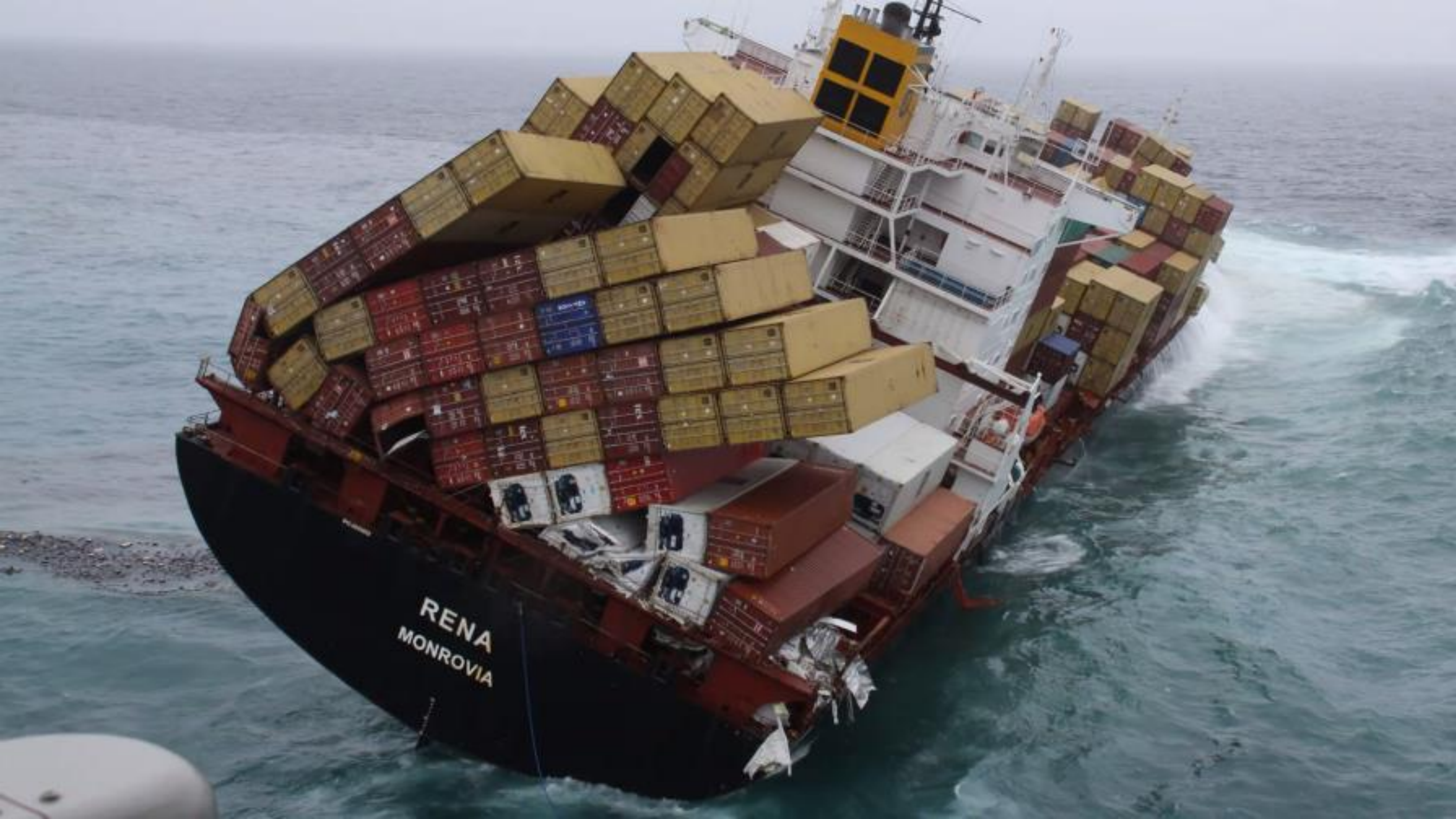
Microsoft

K8S Cluster를 프로덕션 환경에 부드럽게 랜딩 시키기

Sailing Smoothly on K8S

Microsoft App Dev CSA

Inhye Park



RENA
MONROVIA

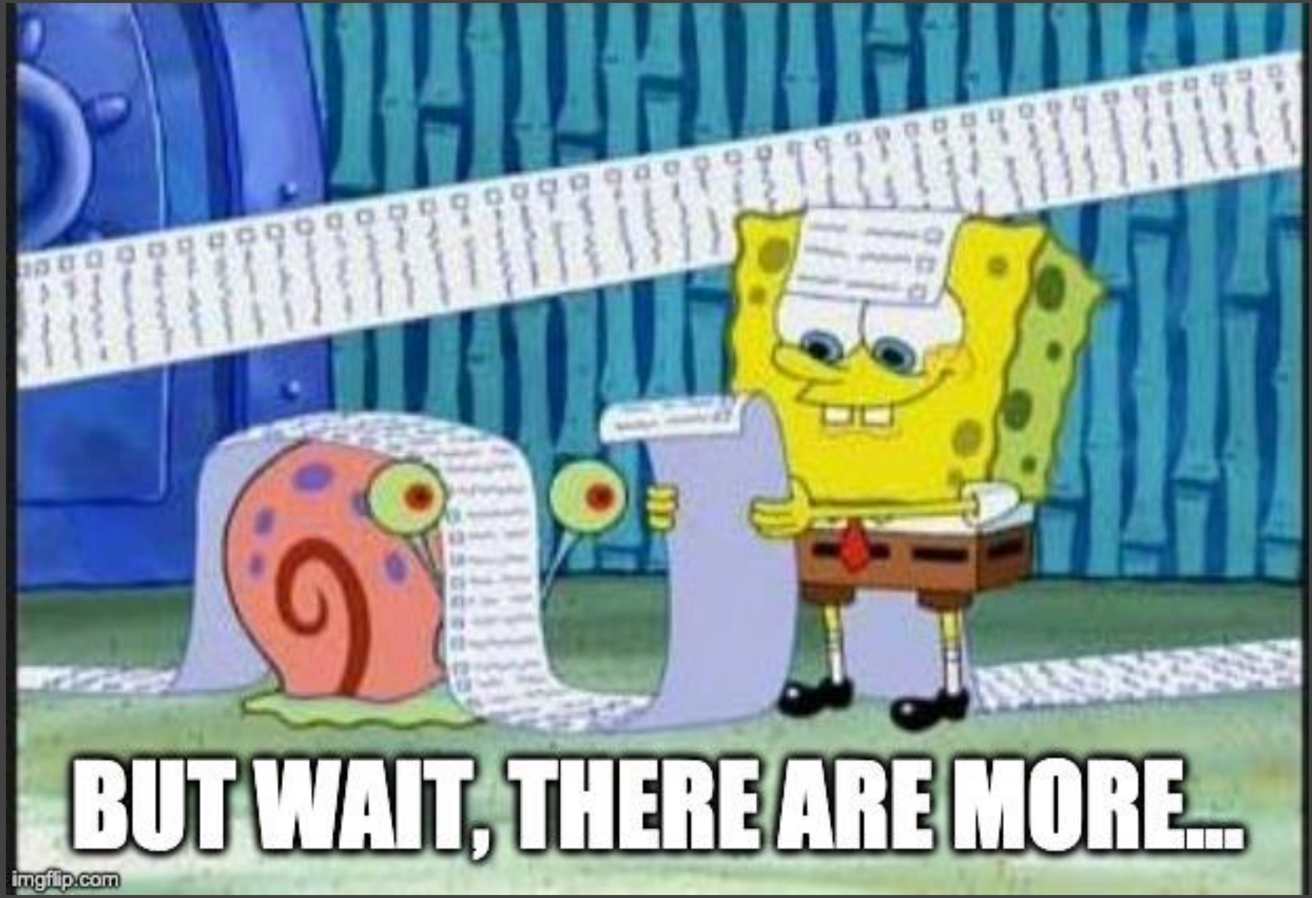
Most common customer issues

Memory overload

IO overload

SNAT port exhaustion

Insufficient quota



- ➢ **Memory overload**
- IO overload
- SNAT port exhaustion
- Insufficient quota

Memory overload

Symptoms

User’s pods get evicted

User’s pods get OOMKilled.

Kube-system pods get evicted.

Nodes get temporarily tainted for not scheduling.

Observations (1)

`No limit is set`

- name: JAVA_OPTS
  value: "**-Xmx2g**"

Describe pod:

Status: Failed

Reason: **Evicted**

Message: The node had condition: [MemoryPressure]

....

Events:

default-scheduler 0/3 nodes are available: 3 node(s) had **taints** that the pod didn't tolerate.

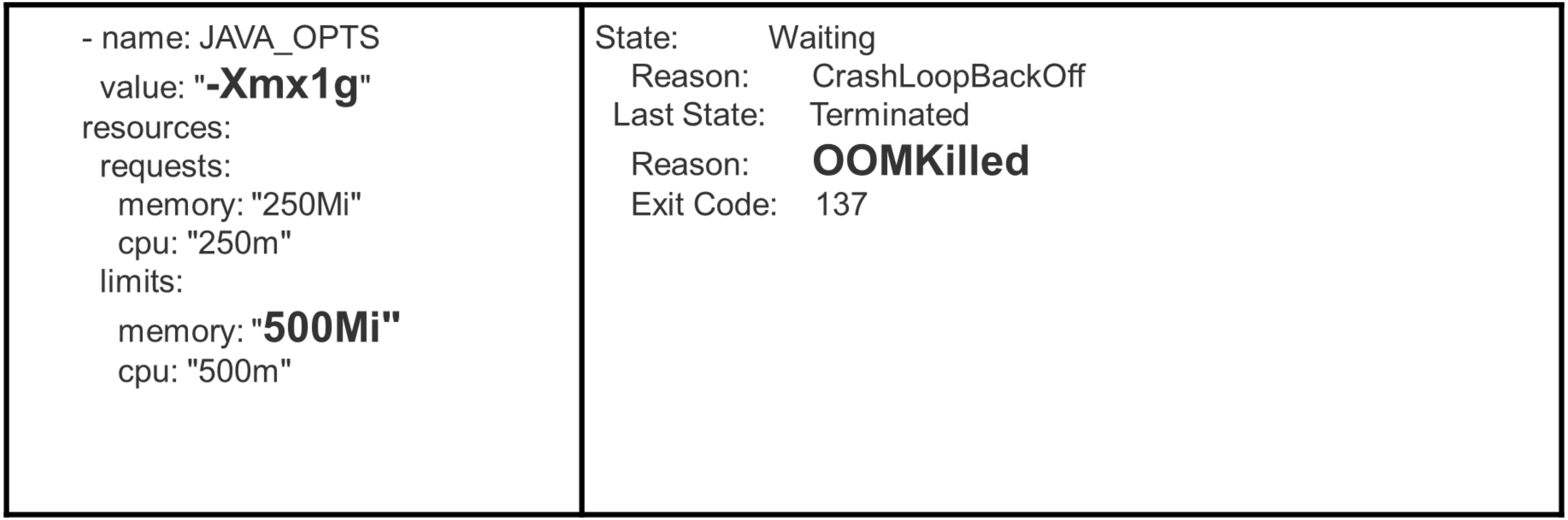
Describe node:

Taints: node.kubernetes.io/**memory-pressure:NoSchedule**

status is now: **NodeHasInsufficientMemory**

Observations (2)

```
Limit is set too low.
```



```
    - name: JAVA_OPTS
      value: "-Xmx1g"
    resources:
      requests:
        memory: "250Mi"
        cpu: "250m"
      limits:
        memory: "500Mi"
        cpu: "500m"
```

```
State:          Waiting
  Reason:       CrashLoopBackOff
 Last State:    Terminated
  Reason:       OOMKilled
  Exit Code:    137
```

Observations (3)

Critical system pods in kube-system namespace get evicted. Cluster goes into unstable or unusable state

```
kube-system   metrics-server-566bd9b4f7-gp9nt        1/1     Running   0          31m
kube-system   metrics-server-566bd9b4f7-zd8lm        0/1     Evicted   0          16h
```

Memory: Best practices

Set resource request and limit on every container.

Set ResourceQuota and/or LimitRange for namespaces.

Isolate system critical pods into their own dedicated node pool.

Put sufficient cores and memory in the pool.

All kube-system pods already have “CriticalAddonsOnly” toleration. Just taint the nodes.

- Memory overload
- ➢ IO overload
- SNAT port exhaustion
- Insufficient quota

IO overload

Symptoms

AKS Cluster nodes going NotReady

“TLS handshake timeout” when reaching the API server.

Critical daemonsets or pods such as kube-proxy, coreDNS start to fail.

Performance and stability issues when using istio or complex operator configurations.

Networking errors or high latency when reaching other Azure services.

Slow DNS queries.

PLEG (pod lifecycle event generator) errors on nodes.

“RPC context deadline exceeded” in kubelet/docker logs.

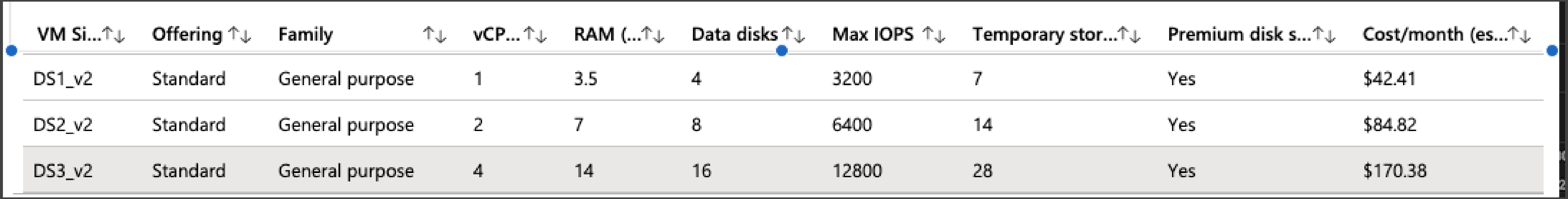
Slow PV attach/detach.

Observations (1)

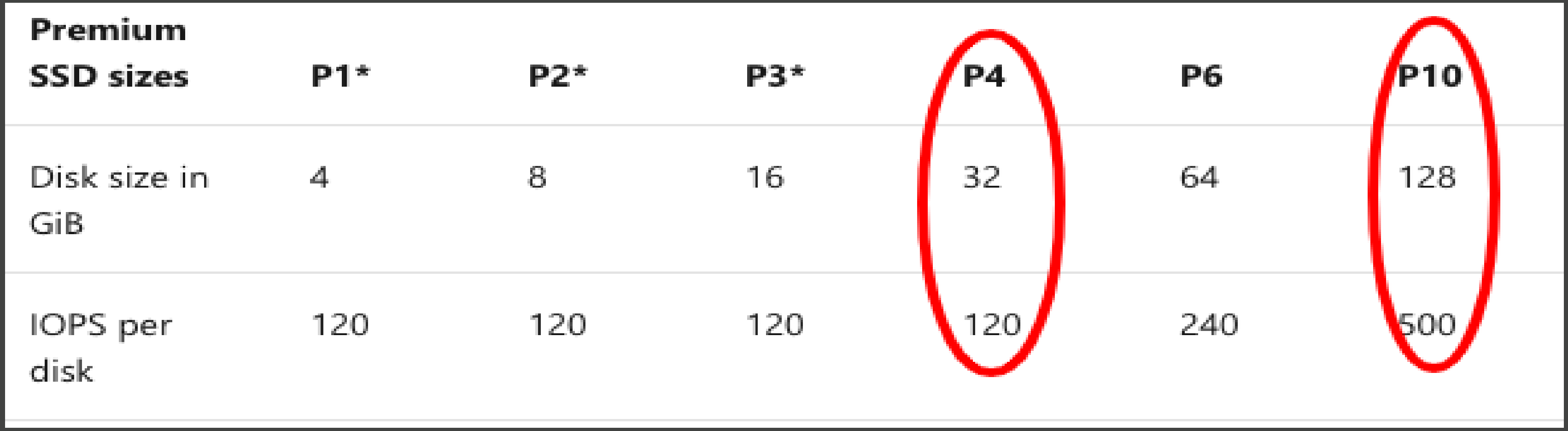
Max IOPS is the lower of VM and the OS disk IOPS.

AKS OS disks are remote disks.

Small OS disk provides low IOPS and throughput.



VM Si...↑↓	Offering ↑↓	Family ↑↓	vCP...↑↓	RAM (...↑↓	Data disks ↑↓	Max IOPS ↑↓	Temporary stor...↑↓	Premium disk s...↑↓	Cost/month (es...↑↓
DS1_v2	Standard	General purpose	1	3.5	4	3200	7	Yes	$42.41
DS2_v2	Standard	General purpose	2	7	8	6400	14	Yes	$84.82
DS3_v2	Standard	General purpose	4	14	16	12800	28	Yes	$170.38



Premium SSD sizes	P1*	P2*	P3*	P4	P6	P10
Disk size in GiB	4	8	16	32	64	128
IOPS per disk	120	120	120	120	240	500

Observations (2)

Usually triggered by:

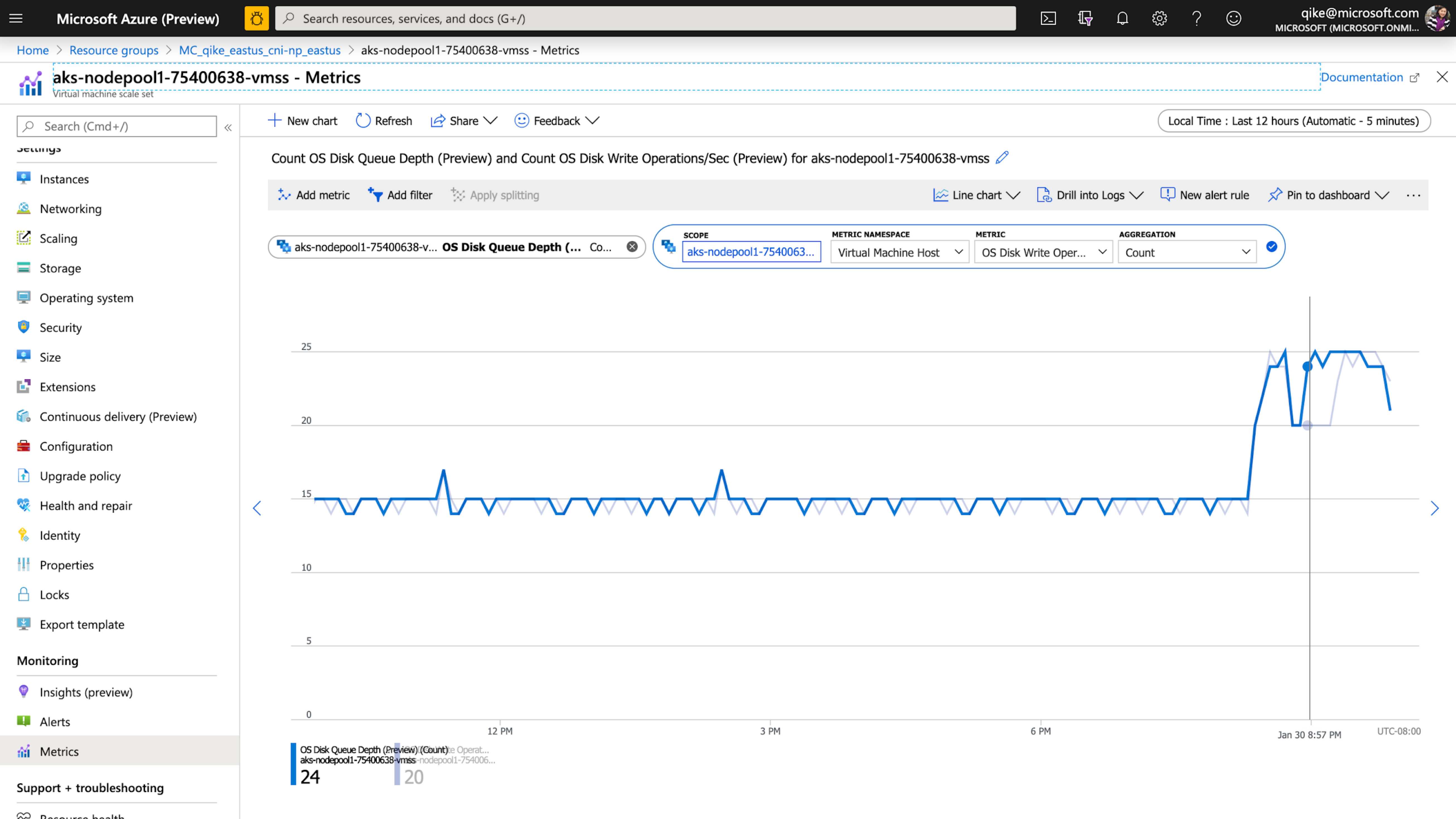
Large number of containers or large container images running on the node.

Aggressive logging (e.g. 3rd party logging, monitoring, audit tools)

Using OS disk as data disk in workload.

Observations (3)

Disk queue depth is an indicator that the OS disk for you worker nodes is throttled.



aks-nodepool1-75400638-vmss - Metrics

Virtual machine scale set

Count OS Disk Queue Depth (Preview) and Count OS Disk Write Operations/Sec (Preview) for aks-nodepool1-75400638-vmss

aks-nodepool1-75400638-v... **OS Disk Queue Depth (...** Co...

SCOPE: aks-nodepool1-7540063...

METRIC NAMESPACE: Virtual Machine Host

METRIC: OS Disk Write Oper...

AGGREGATION: Count

OS Disk Queue Depth (Preview) (Count)
aks-nodepool1-75400638-vmss
24

OS Disk Write Operat...
aks-nodepool1-754006...
20

IO: Best practices

Do not use OS disk for data, use Persistent Volume instead.

Use a sufficiently-sized OS disk.

Use Ephemeral OS Disk (once it is out).

Use knode to mount docker data-drive to ephemeral disk.

Audit I/O from 3rd party add-ons such as Splunk, logstash, filesystem scanners and container scanners.

Set alert for the OS disks

More fun details here: https://aka.ms/aks/io-throttle-issue

- Memory overload
- IO overload

➢ SNAT Port exhaustion

- Insufficient quota

SNAT Port Exhaustion

Symptoms

Intermittent DNS resolution failure.

Node appears NotReady due to unable to reach API server.

Pods get “i/o timeout” when reaching API server or other network addresses.

```
E0124 10:08:30.169432       1 reflector.go:134] github.com/coredns/coredns/plugin/kubernetes/controller.go:317: Failed to list *v1.Endpoints: Get https://xxxx.hcp.eastus.azmk8s.io:443/api/v1/endpoints?limit=500&resourceVersion=0: dial tcp 20.44.xx.xx:443: i/o timeout
```

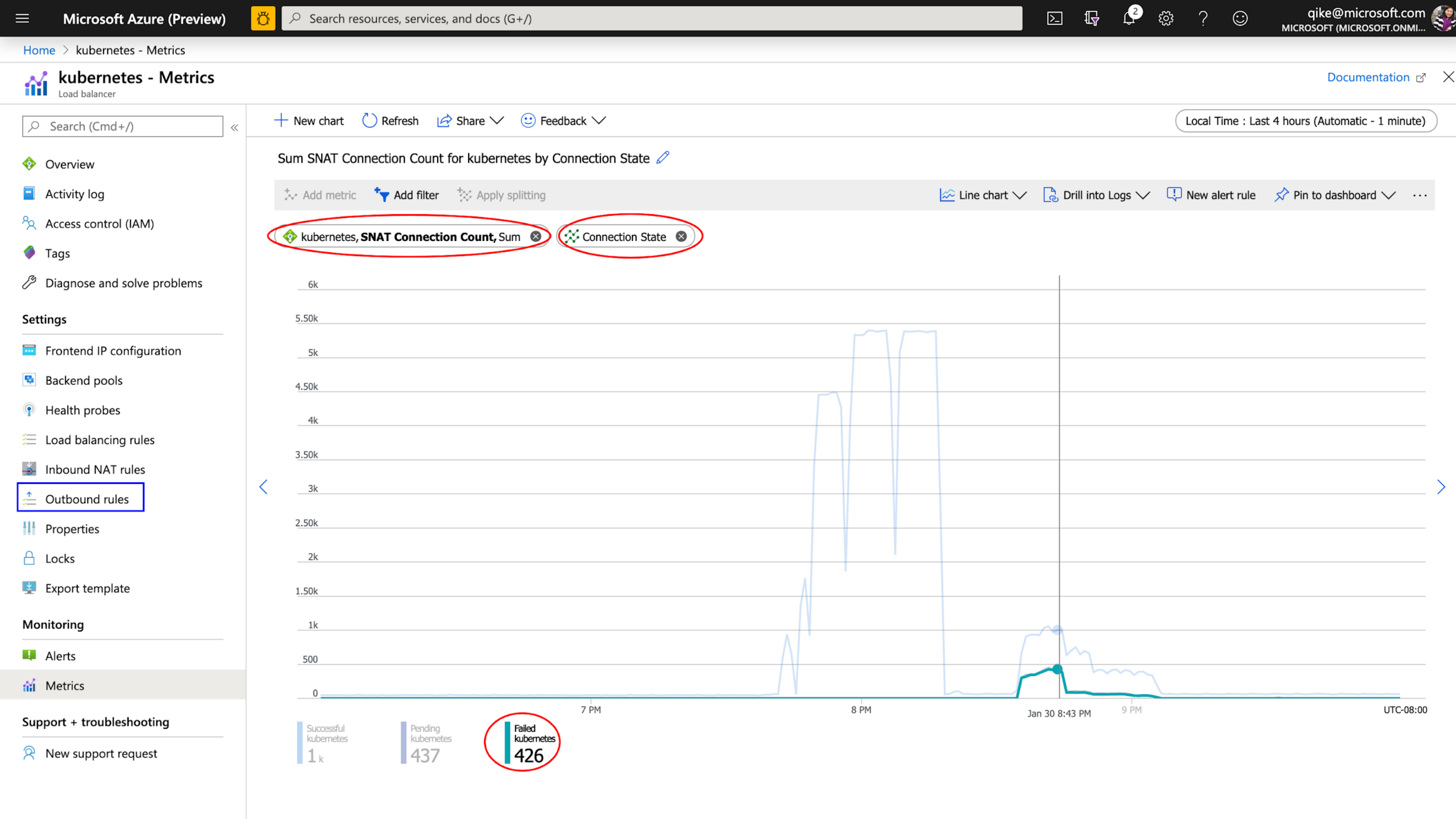
SNAT port exhaustion: Solution

Increase frontend IP.

Increase per VM outbound ports.

Monitor failed “SNAT connections”

Debug using “Used SNAT Ports”, split by “Backend IP Addresses”



kubernetes - Metrics

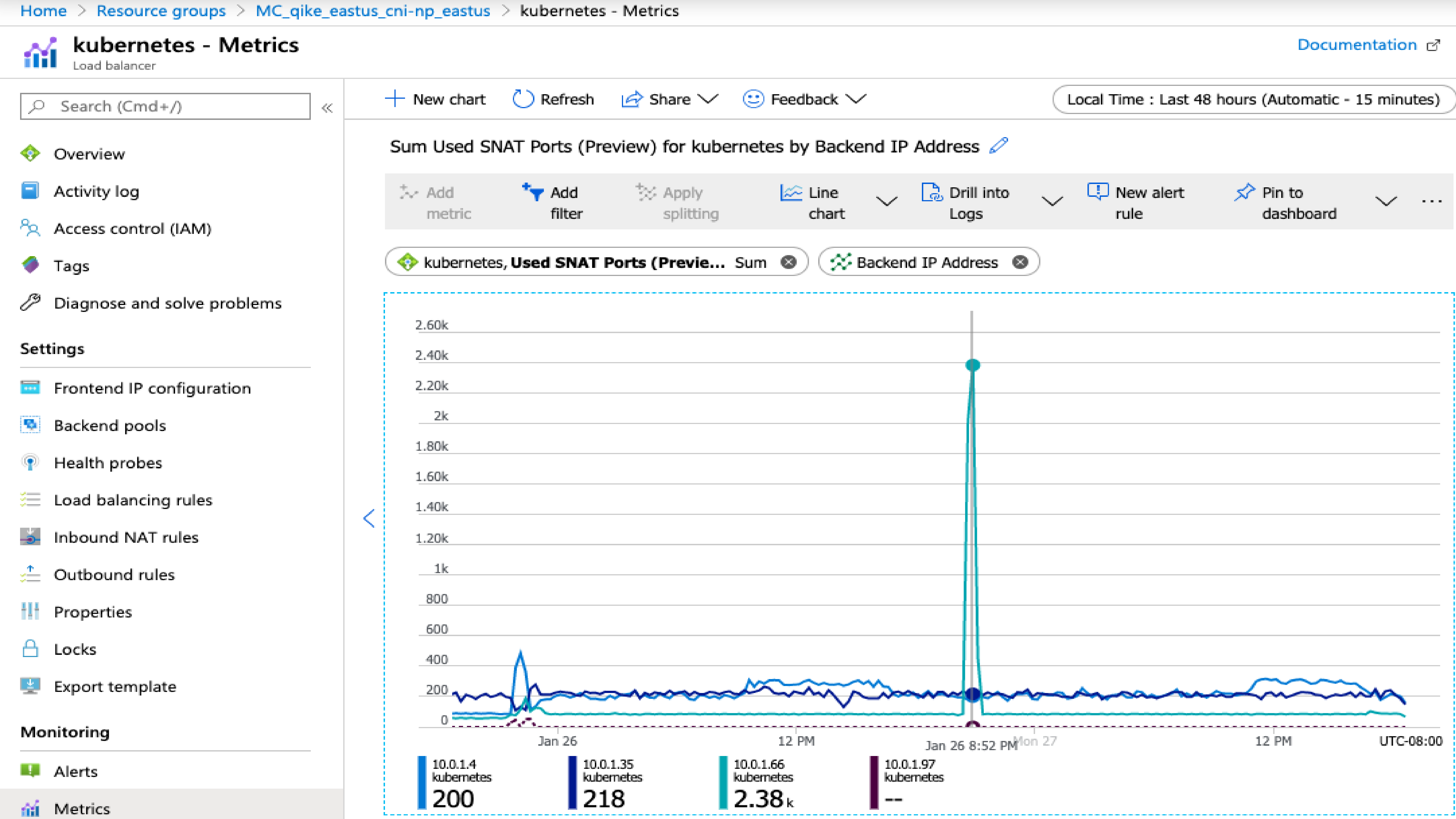
Load balancer

Local Time : Last 4 hours (Automatic - 1 minute)

Sum SNAT Connection Count for kubernetes by Connection State

kubernetes, **SNAT Connection Count,** Sum

Connection State



Home > Resource groups > MC_qike_eastus_cni-np_eastus > kubernetes - Metrics

kubernetes - Metrics

Load balancer

Local Time : Last 48 hours (Automatic - 15 minutes)

Sum Used SNAT Ports (Preview) for kubernetes by Backend IP Address

kubernetes, **Used SNAT Ports (Previe...** Sum

Backend IP Address

10.0.1.4 kubernetes	10.0.1.35 kubernetes	10.0.1.66 kubernetes	10.0.1.97 kubernetes
200	218	2.38 k	--

- Memory overload
- IO overload
- SNAT Port exhaustion

➢ **Insufficient quota**

Insufficient quota

Symptoms

Cluster auto-scaling fails

Manual scaling fails

Upgrade fails

Observations

Cluster upgrade requires at least one more VM (with CPU, GPU, IP).

Using Azure CNI requires additional IPs from the subnet. For each additional node, maxPod + 1 IP addresses are needed.

Higher reliability can cost more, such as using blue/green deployment.

```
$ az aks scale -g qike_rg -n cni-cluster -c 10

Deployment failed. Correlation ID: 98d0c9a1-edb0-414b-9518-xxxxxxx. VMSSAgentPoolReconciler retry
failed: Code="SubnetIsFull" Message="Subnet subnet_cni_2 with address prefix 10.0.1.0/24 does not
have enough capacity for 155 IP addresses."
```

Quota: Best practices

Plan ahead.

Request quota in advance. Sometimes quota can be hard to grant when the inventory in the particular region is low.

Architect the service to run in multiple regions and easy to migrate.

Microsoft