애저 쿠버네티스 서비스 실전 201

Microsoft App Dev CSA

Inhye Park

AKS Compute 심화

Microsoft App Dev CSA

Inhye Park

AKS Compute 심화 - 2

Microsoft App Dev CSA

Inhye Park

AKS Network 심화 - 1

Microsoft App Dev CSA

Inhye Park

AKS Network 심화 - 2

Microsoft App Dev CSA

Inhye Park

AKS Volume 심화

Microsoft App Dev CSA

Inhye Park

K8S Cluster를 프로덕션 환경에 부드럽게 랜딩 시키기

Sailing Smoothly on K8S

Microsoft App Dev CSA

Inhye Park

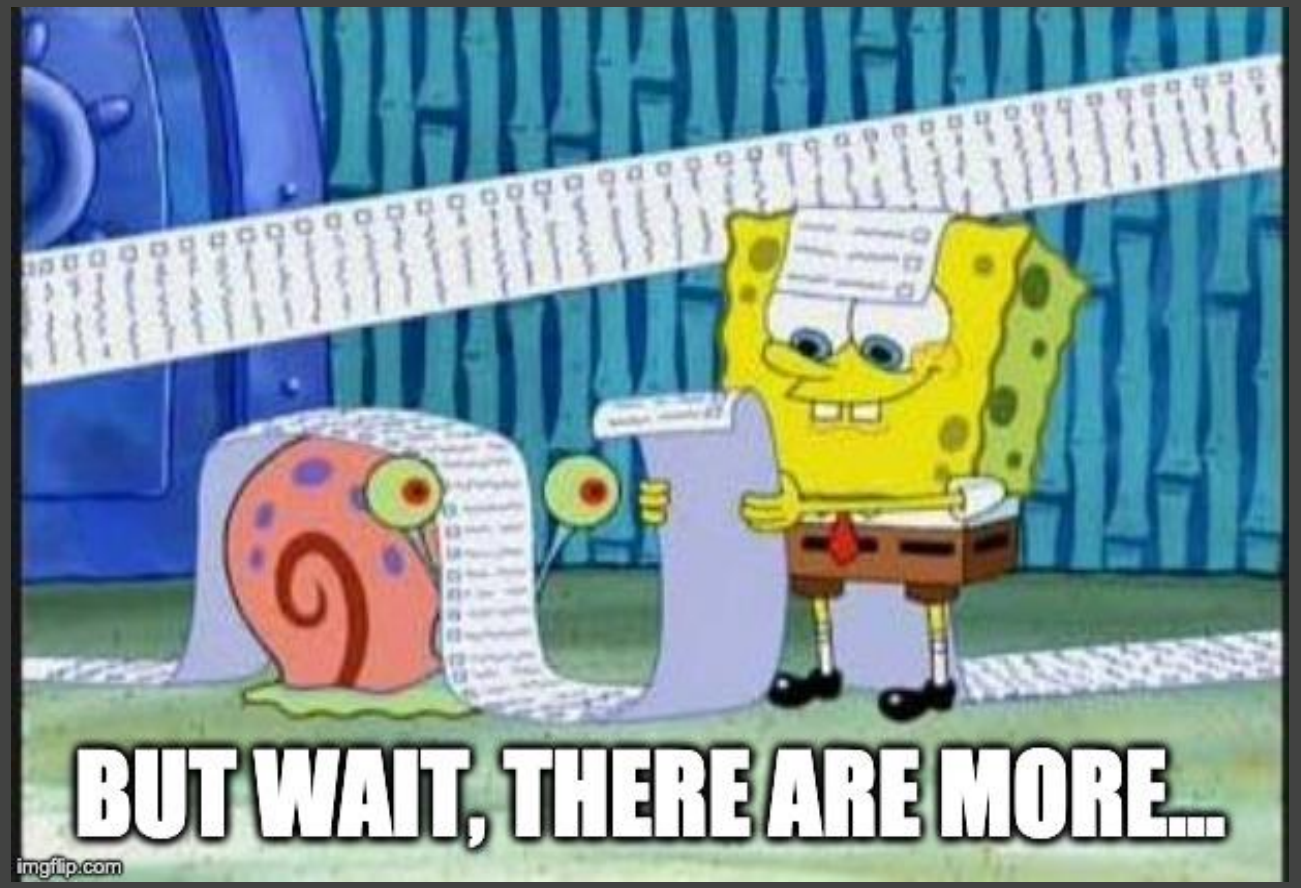
2020/11/12

가장 흔하게 겪는 고객 이슈

메모리 overload

IO overload

SNAT 포트 고갈 & 쿼터 부족



메모리 Overload

- ➢ 메모리 overload
- IO overload
- SNAT 포트 고갈 & 쿼터 부족

증상

유저 Pod가 evict 됨

유저 Pod가 OOMKilled 됨

Kube-system Pod가 evict 됨

노드가 순간적으로 taint되어 스케줄링 할 수 없음

관측 (1)

`Limit` 이 설정되어 있지 않음

- 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**

관측 (2)

Limit은 설정되어 있는데 너무 낮게 설정되어 있을 때.

```
- 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
```

관측 (3)

중요한 kube-system namespace 의 system pod가 eviction 되는 현상. 클러스터가 순간적으로 unstable/ unusable 상태가 될 수 있음

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

메모리: Best practices

모든 컨테이너마다 resource request 와 limit 을 설정해라

네임스페이스에 ResourceQuota와 LimitRange를 설정해라

System critical pod는 dedicated node pool로 Isolation 시켜라

충분한 core와 memory 여유분을 준비해 놓자

모든 kube-system pods는 이미 “CriticalAddonsOnly” 로 toleration 되어 있다.

IO overload

- 메모리 overload
- ➢ IO overload
- SNAT 포트 고갈 & 쿼터 부족

증상

AKS 클러스터 노드들의 상태가 NotReady로 빠짐

API Server와 통신할 때 “TLS handshake timeout” 에러 발생

Kube-proxy, coreDNS 같은 중요한 Daemonsets 또는 pods가 fail 상태가 됨

istio 혹은 복잡한 operator configuration 사용시 퍼포먼스와 안정성에 이슈가 발생됨

다른 Azure Service 와 통신할 때 네트워킹 에러 혹은 높은 latency 로 성능 저하가 발생됨

DNS 쿼리가 늦어짐

노드에 PLEG (pod lifecycle event generator) 에러가 발생됨

kubelet/docker 로그에 “RPC context deadline exceeded” 에러가 발생함

PV attach/detach 가 늦게 됨

관측 (1)

Max IOPS는 VM과 OS disk IOPS 중 낮은 것을 따른다.

AKS OS disks는 리모트 디스크

Small OS disk는 낮은 IOPS와 throughput를 제공함

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

관측 (2)

Usually triggered by:

노드에서 운영되어지고 있는 Containers가 매우 많거나 Container image 사이즈가 클 때

로그 기능을 Aggressive하게 사용 할 경우 (e.g. 3rd party 로깅 시스템, 모니터링, audit tools 사용시)

OS disk를 data disk처럼 사용할 때(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

Documentation

Search (Cmd+/)

Settings

- Instances
- Networking
- Scaling
- Storage
- Operating system
- Security
- Size
- Extensions
- Continuous delivery (Preview)
- Configuration
- Upgrade policy
- Health and repair
- Identity
- Properties
- Locks
- Export template

Monitoring

- Insights (preview)
- Alerts
- Metrics

Support + troubleshooting

- Resource health

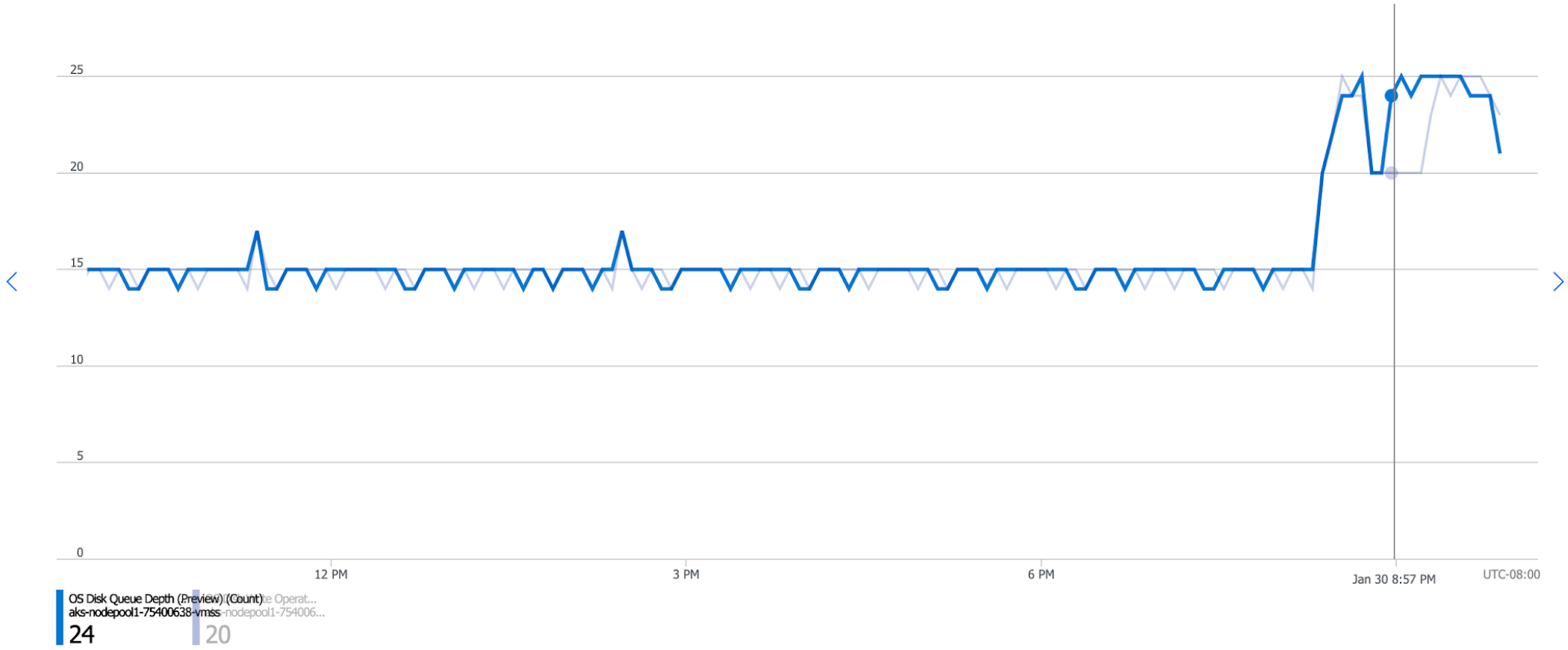
New chart | Refresh | Share | Feedback | Local Time : Last 12 hours (Automatic - 5 minutes)

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

Add metric | Add filter | Apply splitting | Line chart | Drill into Logs | New alert rule | Pin to dashboard | ...

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

SCOPE	METRIC NAMESPACE	METRIC	AGGREGATION
aks-nodepool1-7540063...	Virtual Machine Host	OS Disk Write Oper...	Count



IO: Best practices

OS disk를 data disk처럼 사용하지 말기. 대신에, Persistent Volume을 사용해라.

충분한 사이즈의 OS disk를 사용해야.

Ephemeral OS Disk를 사용해라 (once it is out).

Knode를 사용하여 docker data-drive를 ephemeral disk에 mount 하여 사용해라.

Splunk, logstash, filesystem scanners 그리고 container scanners와 같은 3rd party add-ons 을 사용할 때 I/O를 모니터링 해라

OS disks의 모니터링 값에 alert 을 추가해라

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

SNAT 포트 고갈

- 메모리 overload
- IO overload

➢ SNAT 포트 고갈 & 쿼터 부족

Symptoms

간헐적인 DNS resolution 실패.

API 서버에 연결할 수 없어 노드가 NotReady로 표시됨

Pod가 API server 혹은 other network addresses와 통신할 때 “i/o timeout”에 해당하는 에러를 얻게 됨

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 포트 고갈: 해결책

frontend IP를 증가해라.

VM마다의 outbound ports를 증가시켜라.

"SNAT connection" 의 failed 항목을 모니터링해라

"Backend IP Addresses"로 디바이드하여 "Used SNAT Ports" 메트릭 항목을 모니터링해라

kubernetes - Metrics

Load balancer

Documentation

Search (Cmd+/)

Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems

Settings

Frontend IP configuration
Backend pools
Health probes
Load balancing rules
Inbound NAT rules
Outbound rules
Properties
Locks
Export template

Monitoring

Alerts
Metrics

Support + troubleshooting

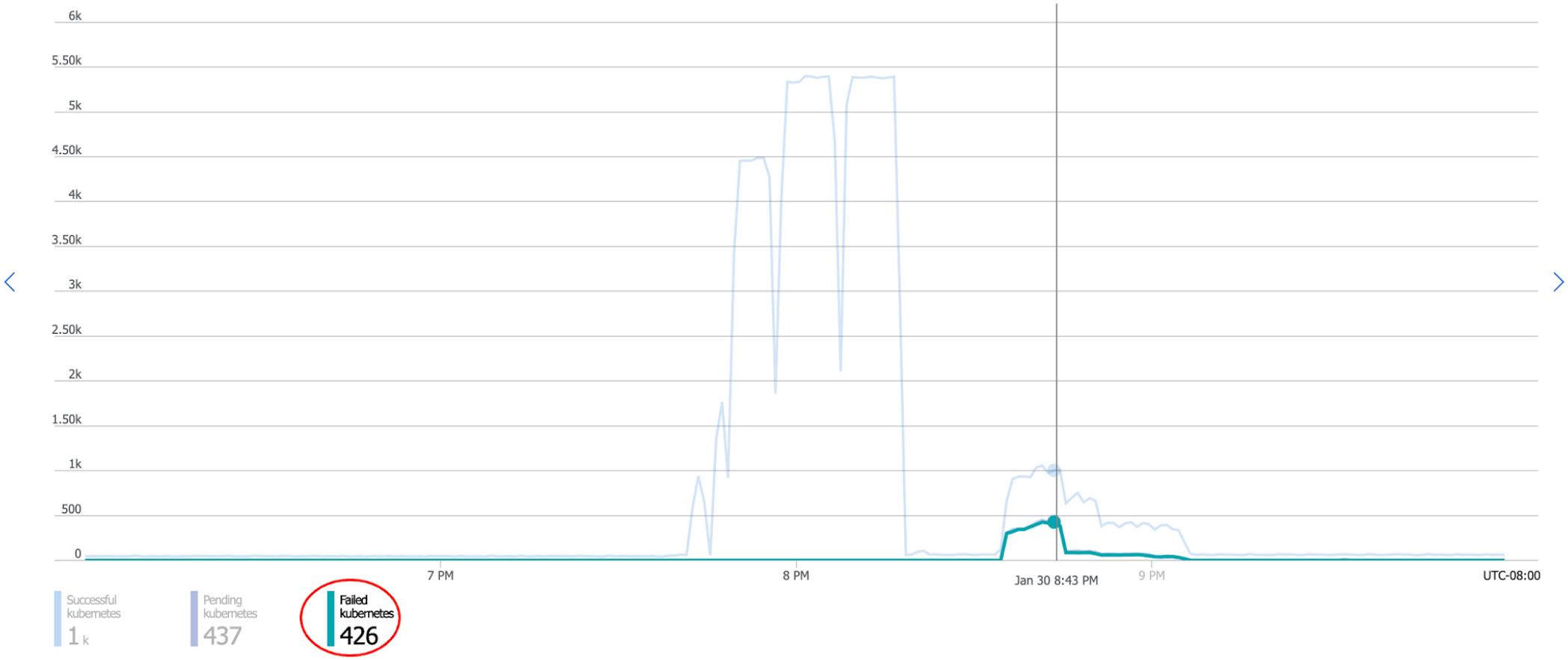
New support request

New chart | Refresh | Share | Feedback | Local Time : Last 4 hours (Automatic - 1 minute)

Sum SNAT Connection Count for kubernetes by Connection State

Add metric | Add filter | Apply splitting | Line chart | Drill into Logs | New alert rule | Pin to dashboard | ...

kubernetes, **SNAT Connection Count**, Sum | Connection State



kubernetes - Metrics

Load balancer

Documentation

Search (Cmd+/)

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems

Settings

- Frontend IP configuration
- Backend pools
- Health probes
- Load balancing rules
- Inbound NAT rules
- Outbound rules
- Properties
- Locks
- Export template

Monitoring

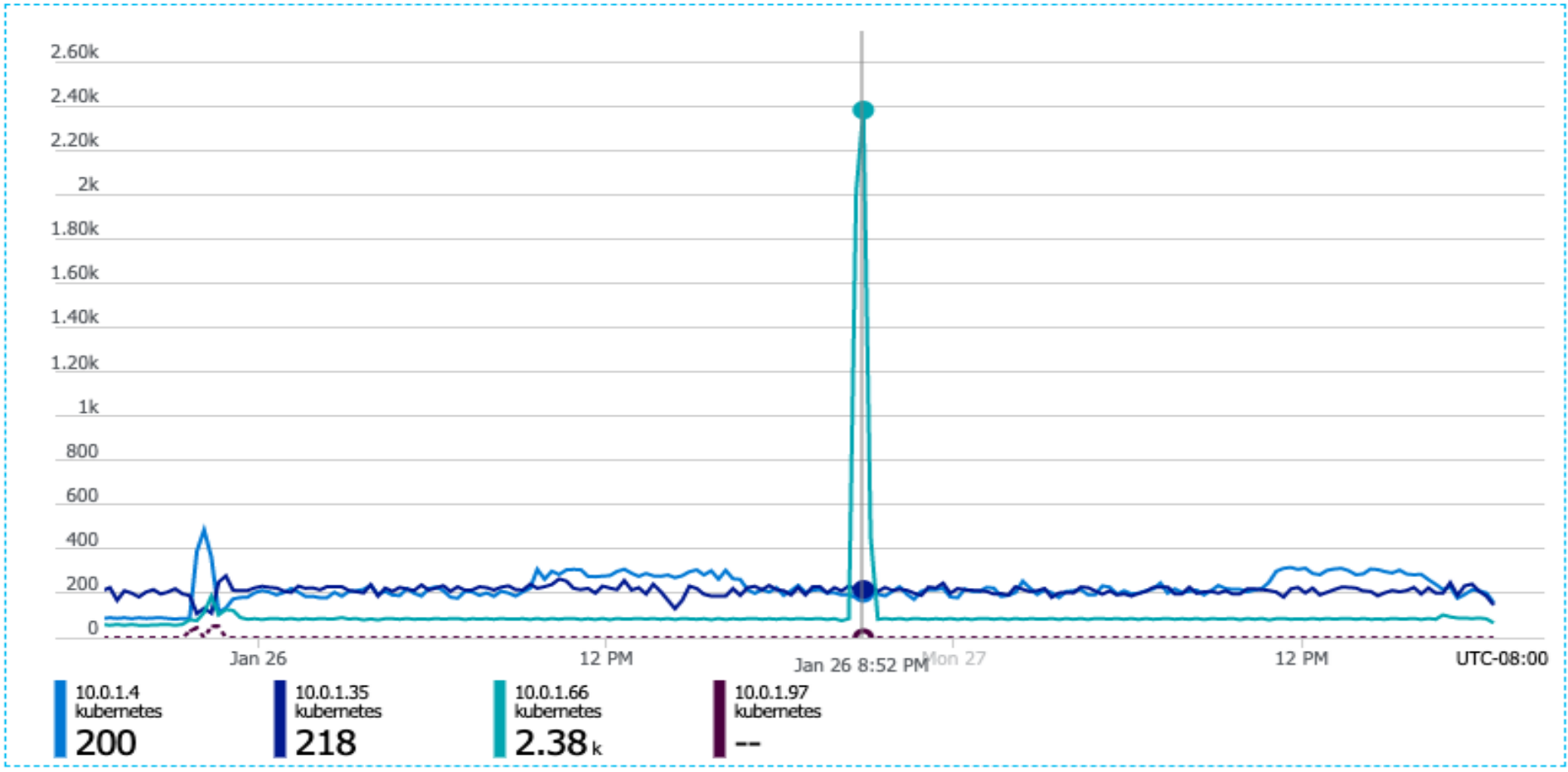
- Alerts
- Metrics

New chart | Refresh | Share | Feedback | Local Time : Last 48 hours (Automatic - 15 minutes)

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

Add metric | Add filter | Apply splitting | Line chart | Drill into Logs | New alert rule | Pin to dashboard | ...

kubernetes, **Used SNAT Ports (Previe...** Sum | Backend IP Address



쿼터 부족

- 메모리 overload
- IO overload
- SNAT 포트 고갈 & 쿼터 부족

증상

Cluster auto-scaling fails

Manual scaling fails

Upgrade fails

관측

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.