Best practises for resources

Tomáš Kukrál

2022-03-30

Motivation -> configuration -> failure

Setting the resources

- Resources per container

- Requests - reservation
- Limits - limitation

- Extended resources
- limits >= requests

```
resources:
  requests:
    cpu: 100m
    memory: 128Mi
    ephemeral-storage: 256Mi
  limits:
    memory: 256Mi
    ephemeral-storage: 256Mi
```

Node resource handling

`kubectl get no -o yaml`

```
allocatable:
  attachable-volumes-gce-pd: "127"
  cpu: 15890m
  ephemeral-storage: "301982796416"
  hugepages-1Gi: "0"
  hugepages-2Mi: "0"
  memory: 56183056Ki
  pods: "110"
capacity:
  attachable-volumes-gce-pd: "127"
  cpu: "16"
  ephemeral-storage: 385926528Ki
  hugepages-1Gi: "0"
  hugepages-2Mi: "0"
  memory: 61739280Ki
  pods: "110"
```

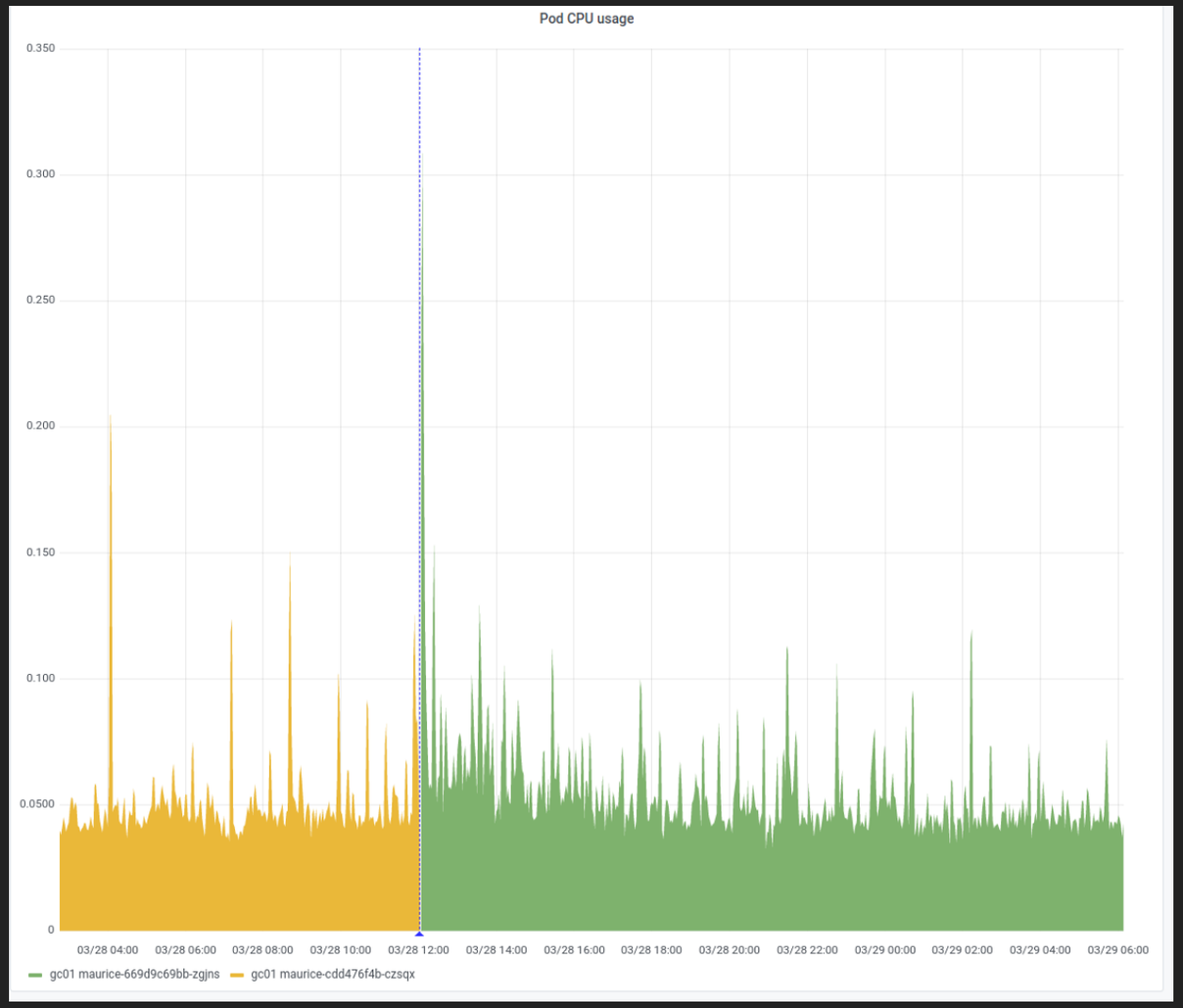
```
kubectl describe no
```

```
  Namespace      Name                          CPU Requests  CPU Limits  Memory Requests  Memory Limits  Age
  ---------      ----                          ------------  ----------  ---------------  -------------  ---
  argo           argocd-dex-server-559ccb5c6   30m (0%)      100m (0%)   32Mi (0%)        64Mi (0%)      33h
  buildbarn      browser-5ffc9dcb76-t4zcx      300m (1%)     500m (3%)   128Mi (0%)       256Mi (0%)     33h
  buildbarn      frontend-f8fb8cf84-tts5f      300m (1%)     500m (3%)   128Mi (0%)       256Mi (0%)     33h
  gitlab         runner-5rpgfcwh-project-580   0 (0%)        0 (0%)      0 (0%)           0 (0%)         65s
  gitlab         runner-nwucgftk-project-580   0 (0%)        0 (0%)      0 (0%)           0 (0%)         65s
  infra          node-config-rxfjr             20m (0%)      150m (0%)   64Mi (0%)        180Mi (0%)     34h
  kube-system    anetd-slvf2                   100m (0%)     0 (0%)      100Mi (0%)       0 (0%)         34h
  kube-system    gke-metadata-server-ppt29     100m (0%)     100m (0%)   100Mi (0%)       100Mi (0%)     34h
  kube-system    netd-d22vg                    0 (0%)        0 (0%)      0 (0%)           0 (0%)         34h
  kube-system    pdcsi-node-qdcjh              10m (0%)      0 (0%)      20Mi (0%)        100Mi (0%)     34h
  loki           promtail-n97rv                20m (0%)      150m (0%)   64Mi (0%)        180Mi (0%)     34h
  prometheus     node-exporter-hpwcs           0 (0%)        0 (0%)      0 (0%)           0 (0%)         34h
```

CPU

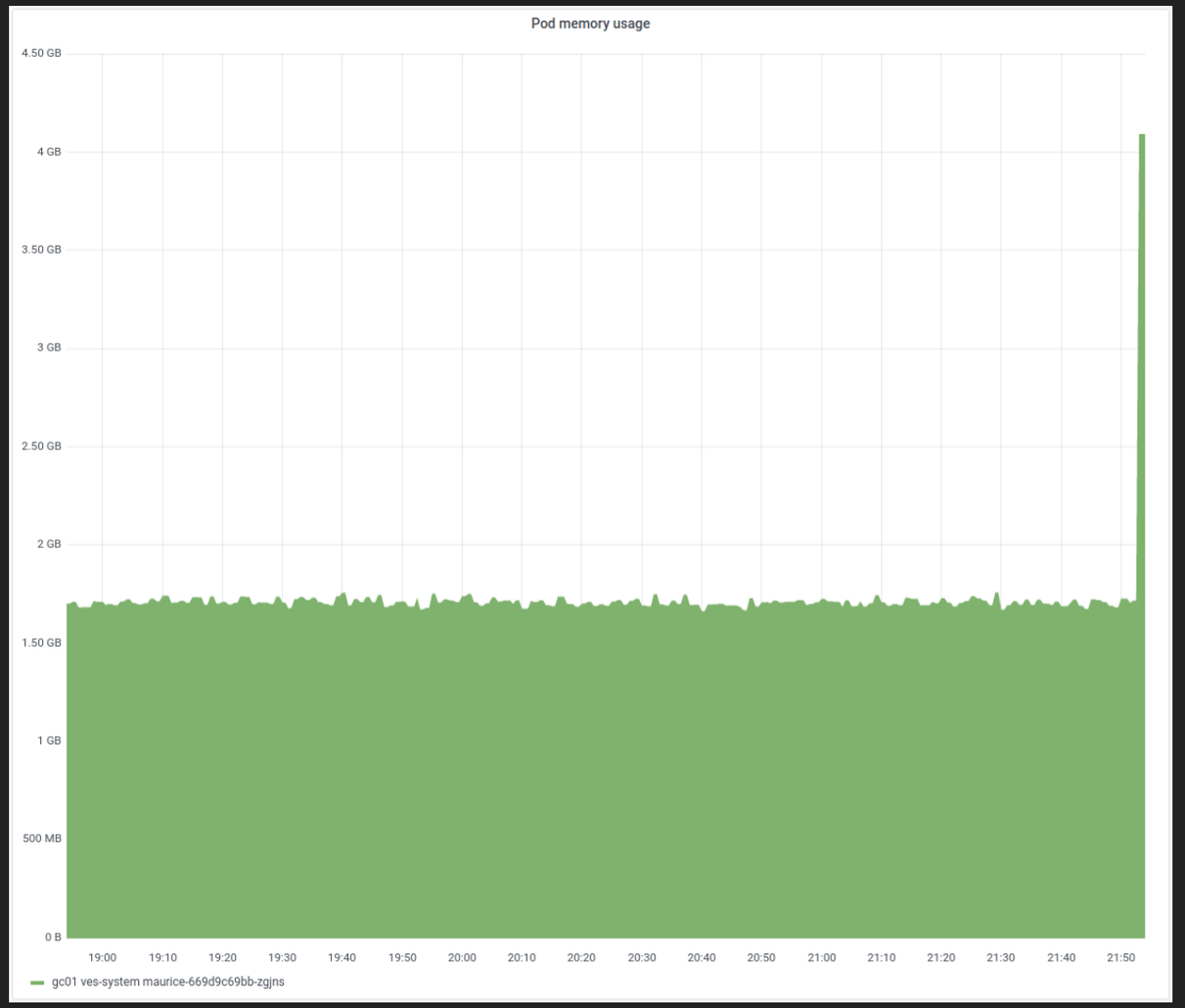
- Throttling
- CPU unit
- Time slide
- Absolute

- `cpu_share`
- `cpu_quota`
- `cpu_period`



Memory

- OOM
- bytes
- EmptyDir with memory medium



Storage

- bytes
- logs
- container write layer
- quota enforcing

ResourceQuota

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: kad-quotas
  namespace: kad
spec:
  hard:
    requests.cpu: 200m
    requests.memory: 256Mi
    limits.cpu: 1
    limits.memory: 512Mi
```

LimitRange

```
apiVersion: v1
kind: LimitRange
metadata:
  name: resource-limits
  namespace: kad
spec:
  limits:
  - min:
      cpu: 50m
  - max:
      cpu: 500m
  - default:
      cpu: 100m
      memory: 16Mi
    defaultRequest:
      cpu: 50m
      memory: 16Mi
    type: Container
```

Pod QoS

- `oom_score_adj`

- Guaranteed `997`
- BestEffort `1000`
- Burstable `memReq/MemAll * 1000`

```
 kubectl get po gitlab-runner-blue-75fdc5947d-zlnms -o json | jq -r .status.qosClass
Burstable
```

Failures

- Eviction
- Pod priority

- Restart vs Pending
- exit 137

Debugging

```
kubectl top pod

NAME                                                CPU(cores)   MEMORY(bytes)
gitlab-runner-blue-75fdc5947d-zlnms                 59m          119Mi
gitlab-runner-brown-6775bcc76b-rh46t                1m           12Mi
gitlab-runner-green-7b8977bd4f-px6tw                100m         64Mi
runner-5rpgfcwh-project-10189624-concurrent-0m5l6j  0m           2Mi
runner-5rpgfcwh-project-10479312-concurrent-0lnjtw  2012m        417Mi
runner-5rpgfcwh-project-10479312-concurrent-3xbskf  2568m        480Mi
runner-5rpgfcwh-project-6351664-concurrent-2nw9t8   988m         52Mi
runner-5rpgfcwh-project-6351664-concurrent-3sz2tg   1482m        1923Mi
runner-nwucgftk-project-10479312-concurrent-0d5tqw  0m           3Mi
runner-nwucgftk-project-13212207-concurrent-08fmdm  128m         799Mi
runner-nwucgftk-project-5988224-concurrent-3l7f6q   537m         27000Mi
runner-nwucgftk-project-6002292-concurrent-2n9t8f   720m         5977Mi
runner-nwucgftk-project-6351664-concurrent-07n7t4   340m         363Mi
runner-nwucgftk-project-6351664-concurrent-1cmfw7   1079m        51Mi
```

```
kubectl describe node

...
Non-terminated Pods:          (14 in total)
  Namespace                   Name                                                  CPU Requests  CPU Limits  Memory Requests  Memory Limits  Age
  ---------                   ----                                                  ------------  ----------  ---------------  -------------  ---
  argo                        argocd-dex-server-559ccb5c6f-wntm7                    30m (0%)      100m (0%)   32Mi (0%)        64Mi (0%)      2d1h
  buildbarn                   browser-5ffc9dcb76-t4zcx                              300m (1%)     500m (3%)   128Mi (0%)       256Mi (0%)     2d1h
  buildbarn                   frontend-f8fb8cf84-tts5f                              300m (1%)     500m (3%)   128Mi (0%)       256Mi (0%)     2d1h
  gitlab                      runner-nwucgftk-project-10479312-concurrent-0d5tqw    0 (0%)        0 (0%)      0 (0%)           0 (0%)         3m10s
  gitlab                      runner-nwucgftk-project-6351664-concurrent-0xxfrm     0 (0%)        0 (0%)      0 (0%)           0 (0%)         98s
  gitlab                      runner-nwucgftk-project-6351664-concurrent-344rmn     0 (0%)        0 (0%)      0 (0%)           0 (0%)         97s
  gitlab                      runner-nwucgftk-project-6351664-concurrent-5vzxng     0 (0%)        0 (0%)      0 (0%)           0 (0%)         74s
  infra                       node-config-rxfjr                                     20m (0%)      150m (0%)   64Mi (0%)        180Mi (0%)     2d3h
  kube-system                 anetd-slvf2                                           100m (0%)     0 (0%)      100Mi (0%)       0 (0%)         2d3h
  kube-system                 gke-metadata-server-ppt29                             100m (0%)     100m (0%)   100Mi (0%)       100Mi (0%)     2d3h
  kube-system                 netd-d22vg                                            0 (0%)        0 (0%)      0 (0%)           0 (0%)         2d3h
  kube-system                 pdcsi-node-qdcjh                                      10m (0%)      0 (0%)      20Mi (0%)        100Mi (0%)     2d3h
  loki                        promtail-n97rv                                        20m (0%)      150m (0%)   64Mi (0%)        180Mi (0%)     2d3h
  prometheus                  node-exporter-hpwcs                                   0 (0%)        0 (0%)      0 (0%)           0 (0%)         2d3h
```

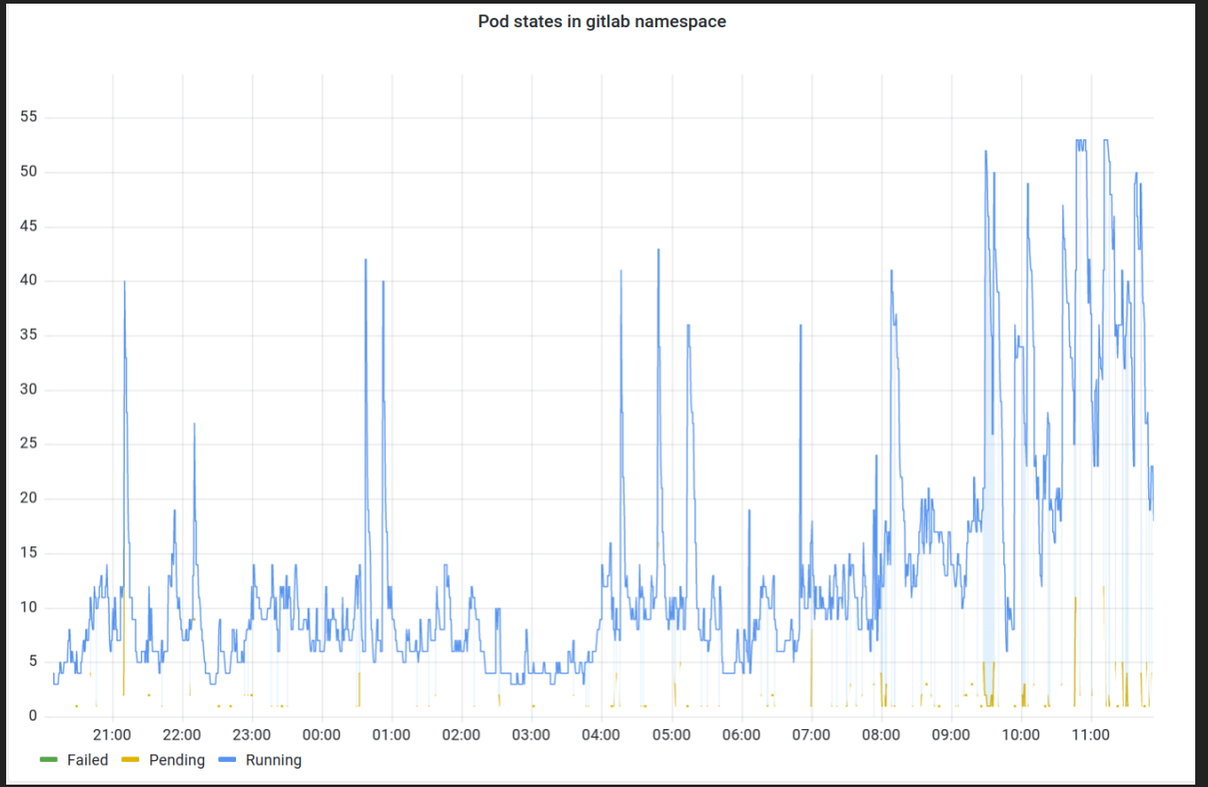
```
kubectl get po --field-selector spec.nodeName=gke-gc01-cf25a10e-nroz
```

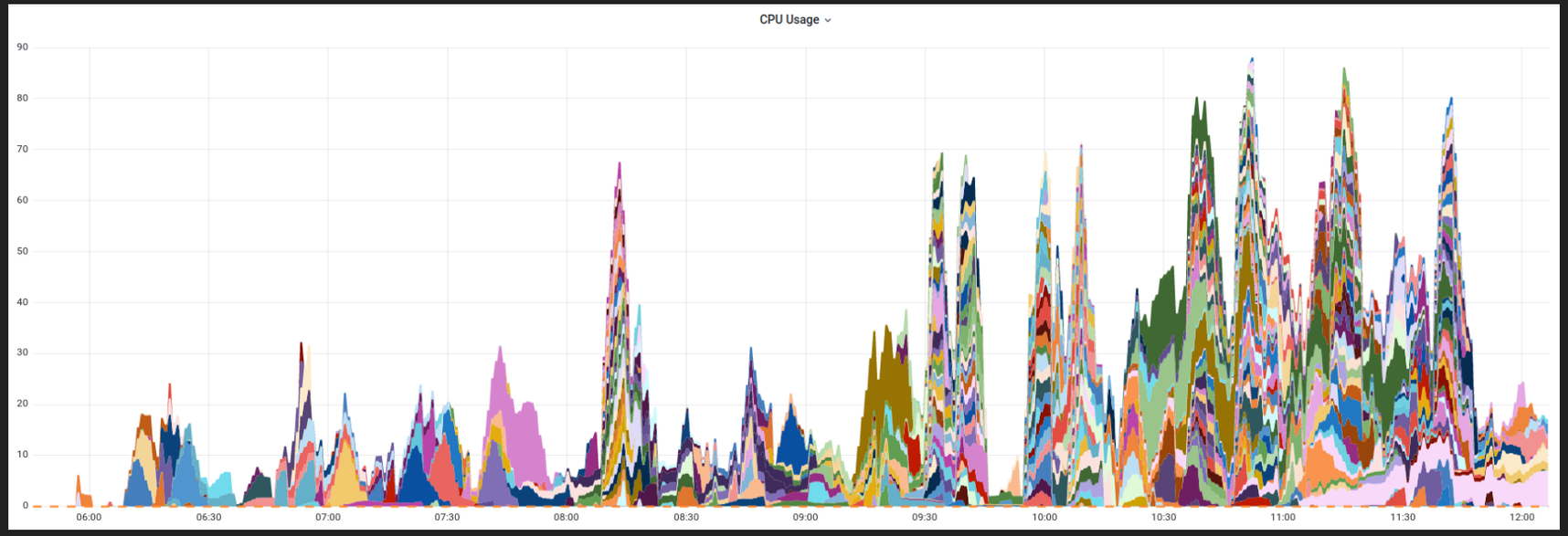
Unpredictable workload

Pods (jobs) with unrestricted resource usage

- 0..20 nodes (500$/node/month)
- 0..19GB memory usage
- any CPU usage
- 0..40GB filesystem usage
- 0..6h lifetime
- various image sizes

- Golang test
- Image builds (podman/buildah)




CPU Usage
90
80
70
60
50
40
30
20
10
0
06:00
06:30
07:00
07:30
08:00
08:30
09:00
09:30
10:00
10:30
11:00
11:30
12:00

- 2 node pools
  - static - 1..4 nodes
  - pool- 0..20
- scaling max jobs according to expected load (6 jobs per node)
- graceful node termination
- automatic scaling
- node rotation

QA