

Kubernetes w eksploatacji

Sondy, zasoby, skalowanie i trwałe magazyny

mgr inż. Jakub Woźniak

Zarządzanie Systemami Rozproszonymi · Semestr zimowy 2026/27

Sonda	Niepowodzenie po progu	Przykład
startup	Restart kontenera	Długa inicjalizacja cache
readiness	Pod niegotowy do zwykłego ruchu Service	API jeszcze nie obsługuje żądań
liveness	Restart kontenera	Trwale zablokowana pętla procesu

- Do sukcesu startup pozostałe sondy są wstrzymane.
- Readiness nie zatrzymuje automatycznie odbioru wiadomości przez workera.

```
ports: [{containerPort: 8080}]
startupProbe:
  httpGet: {path: /startup, port: 8080}
  periodSeconds: 5
  failureThreshold: 24
readinessProbe:
  httpGet: {path: /ready, port: 8080}
  periodSeconds: 5
  failureThreshold: 2
livenessProbe:
  httpGet: {path: /live, port: 8080}
  periodSeconds: 10
  failureThreshold: 3
```

Około 120 s budżetu startup. Rzeczywiste czasy zależą również od harmonogramu sond i timeoutów.

Baza niedostępna



/live sprawdza bazę



Restart wszystkich API

- Restart API nie przywraca bazy; usuwa rozgrzane połączenia i cache.
- Liveness: warunek, którego naprawą rzeczywiście może być restart procesu.
- Readiness: zdolność do obsługi ruchu; zależności sprawdzane świadomie.

API z cache

Część odczytów może działać mimo awarii bazy. Jedna wspólna sonda nie opisuje wszystkich tras.

```
resources:  
  requests: {cpu: "250m", memory: "128Mi"}  
  limits:   {cpu: "500m", memory: "256Mi"}
```

Pole	Znaczenie	Obserwacja
CPU request	Wejście schedulera; udział przy konkurencji	Nie jest limitem 250m
CPU limit	Ograniczanie czasu CPU	Throttling, wzrost opóźnień
Memory limit	Granica pamięci kontenera	Możliwy OOM kill

Uproszczony model

Nowe repliki = zaokrąglenie w górę(aktualne repliki × metryka / cel).

- 3 Pody, każdy request CPU = 200m; średnie użycie = 160m.
- Wykorzystanie względem request = 80%; cel HPA = 50%.
- Nowa liczba: $\text{ceil}(3 \times 80 / 50) = 5$.
- Założenie: kompletne metryki, gotowe Pody, brak innych ograniczeń.

Kubernetes — HPA algorithm details

```
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata: {name: orders-api}
spec:
  scaleTargetRef:
    {apiVersion: apps/v1, kind: Deployment, name: orders-api}
  minReplicas: 3
  maxReplicas: 8
  metrics:
    - type: Resource
      resource:
        name: cpu
        target: {type: Utilization, averageUtilization: 50}
```

Wymaga API metryk. Stabilizacja, tolerancja i brakujące próbki modyfikują prosty rachunek.

PVC: żądanie 10 GiB



StorageClass / CSI



PV: przydzielony wolumen

- PVC: żądanie pojemności i sposobu dostępu przez aplikację.
- PV: obiekt reprezentujący magazyn; może powstać dynamicznie.
- StorageClass: parametry provisioningu, np. sterownik i polityka wiązania.
- Usunięcie PVC: skutek zależy m.in. od reclaimPolicy; możliwe usunięcie danych.



- Kontroler bramy realizuje routing host/path i TLS.
- Service daje stabilny punkt dostępu; endpointy wynikają z selektora i gotowości.
- Przypadek standardowy: ruch kierowany do gotowych endpointów.

Istnieją wyjątki konfiguracji, np. `publishNotReadyAddresses`.

[Gateway API — API overview](#)

Zdarzenie	Pytanie o storage
Nowy Pod	Czy może ponownie zamontować wolumen?
Utrata węzła	Czy wolumen jest niezależny od dysku lokalnego?
Utrata strefy	Czy dane i możliwość montowania istnieją poza strefą?

- Topologia Podów musi być zgodna z topologią wolumenu.
- Trwałość dysku nie zastępuje kopii chroniącej przed błędnym DELETE.

Kubernetes — Storage classes

HPA: 3 → 5



Brak zasobów na węzłach



Dodatkowe Pody Pending

- Autoscaler węzłów może dodać pojemność; podlega limitom i opóźnieniom.
- Więcej workerów może przeciążyć bazę lub wyczerpać limit połączeń.
- Maksymalna skala wymaga testu zależności oraz oszacowania kosztu.

Microsoft — AKS cluster autoscaler

Last State: Terminated

Reason: OOMKilled

Exit Code: 137

Restart Count: 12

Limits: memory: 256Mi

Pomiary tuż przed restartem

RSS: 120 MiB -> 180 MiB -> 254 MiB

CPU: 60m; request CPU: 250m

Decyzja

Zmienić sondę, zwiększyć pamięć czy dodać repliki? Oddziel działanie doraźne od ustalenia przyczyny.

- [Kubernetes — Probes](#)
- [Kubernetes — Resource management](#)
- [Kubernetes — HPA](#)
- [Kubernetes — Persistent Volumes](#)
- [Gateway API — API overview](#)
- [Microsoft — AKS cluster autoscaler](#)

Dokumentacja sprawdzona: 26.09.2026. Dane przypadków: przykłady dydaktyczne.