купер

4 способа выявления сбоев узла Kubernetes: актуальные стратегии возвращения рабочей нагрузки

Дмитрий Рыбалка

Техлид отдела базовой ИТ-инфраструктуры в Купере

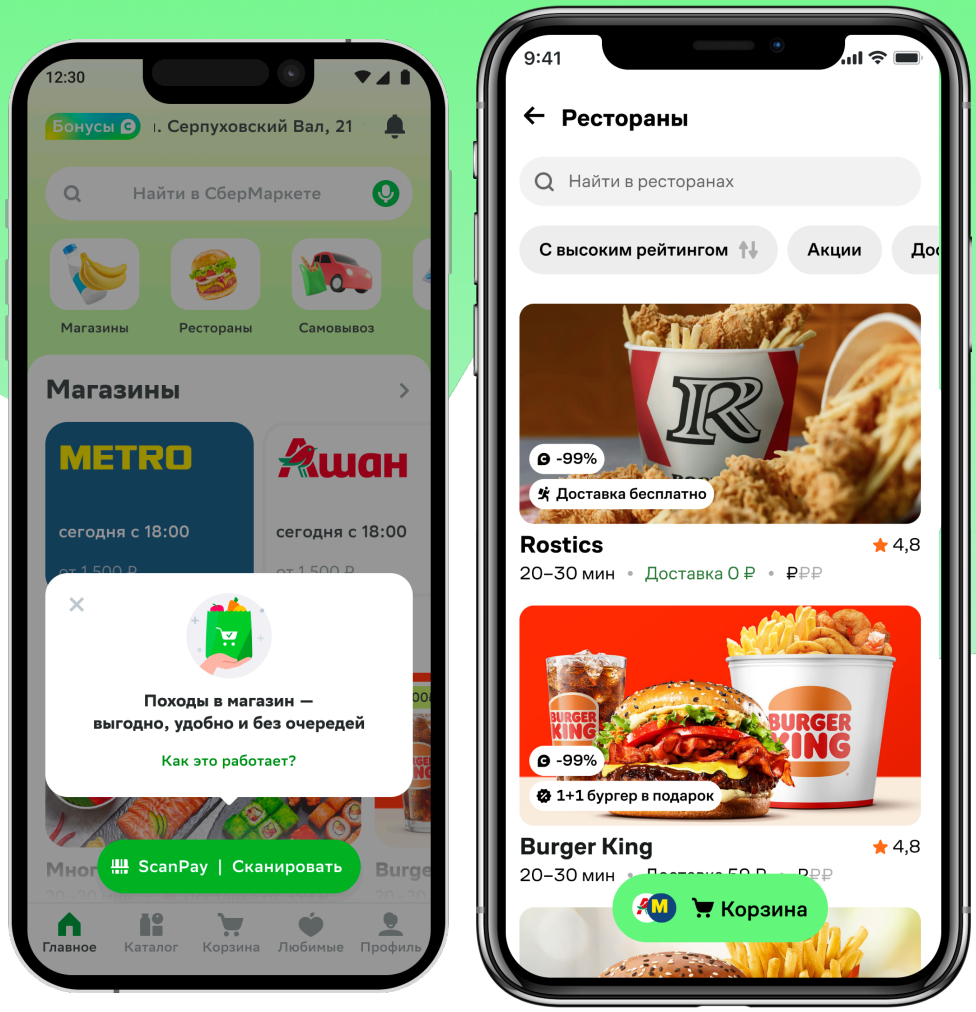
Что видит клиент когда …

Отказывает:

1 узел

10

1 zone



Что происходит внутри...

SLI Kafka

Доступность брокеров kafka

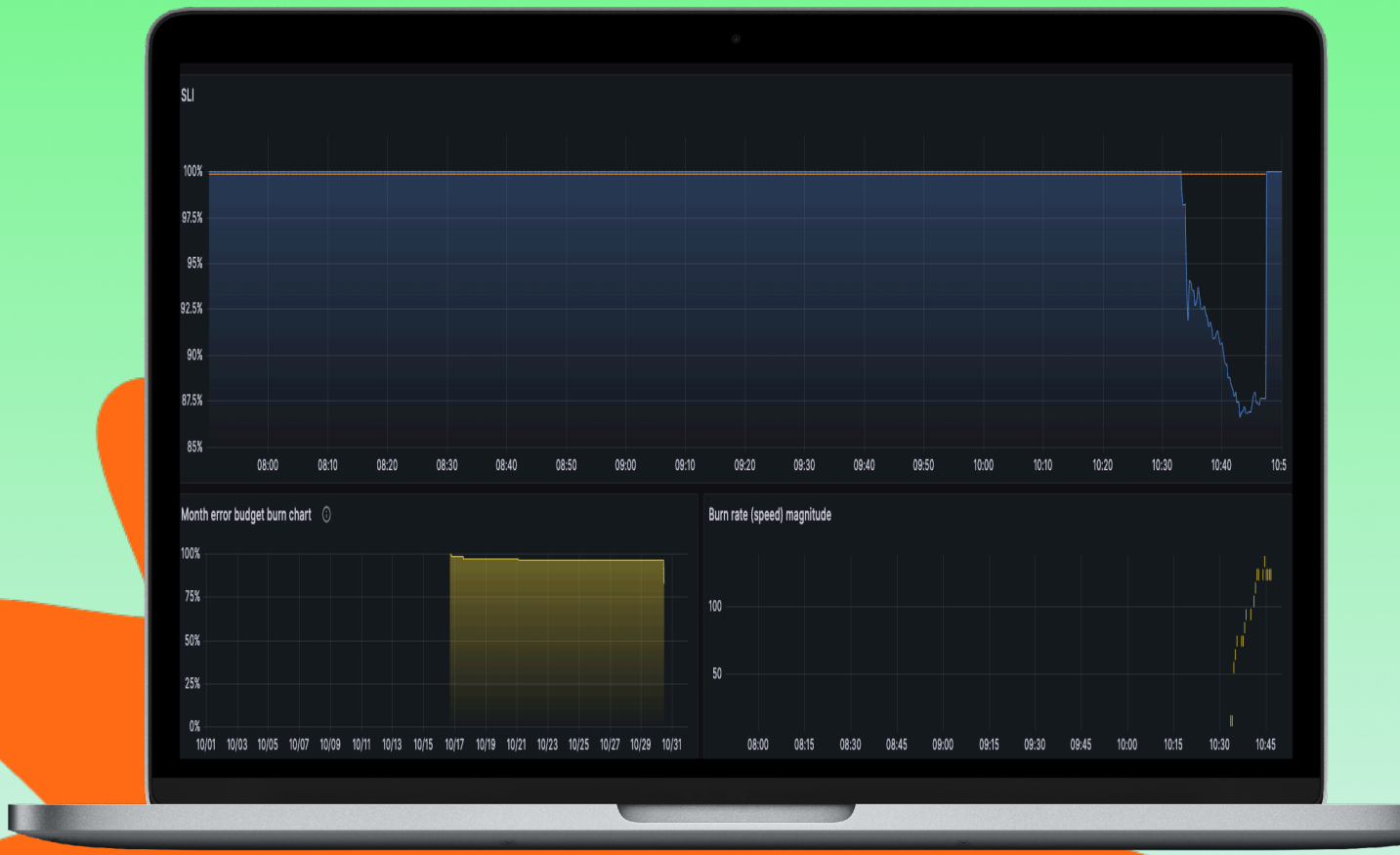
Латенси kafka

Что происходит внутри...

SLI сервиса

Бюджет ошибок горит

Dev команда негодует



О компании

tech

СБЕР МАРКЕТ

Cloud Agnostic / Kubernetes first

Kubernetes

1 Регион

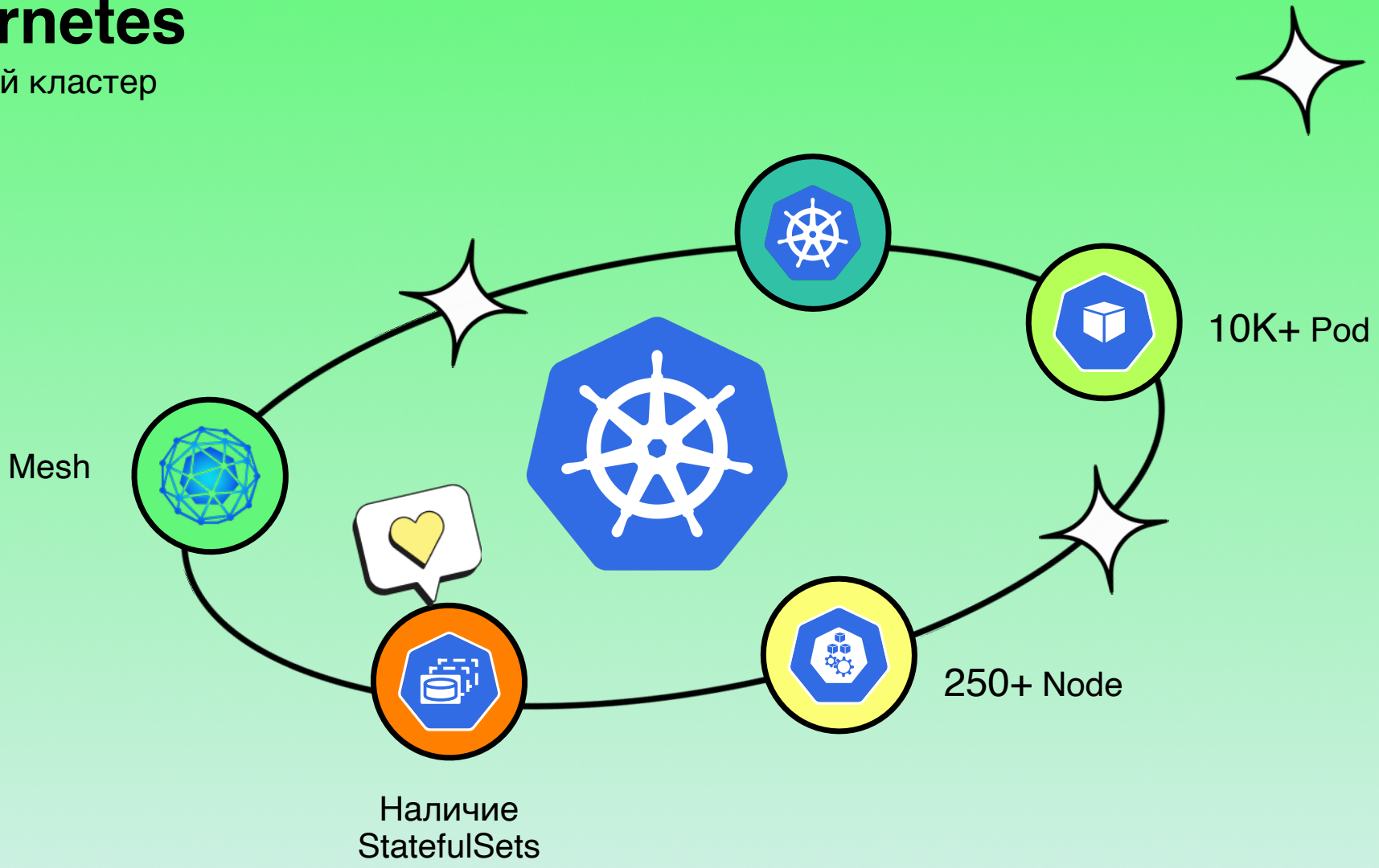
5 Зон доступности

15+ Prod Кластеров

Размещены в cloud

Kubernetes

Типовой кластер

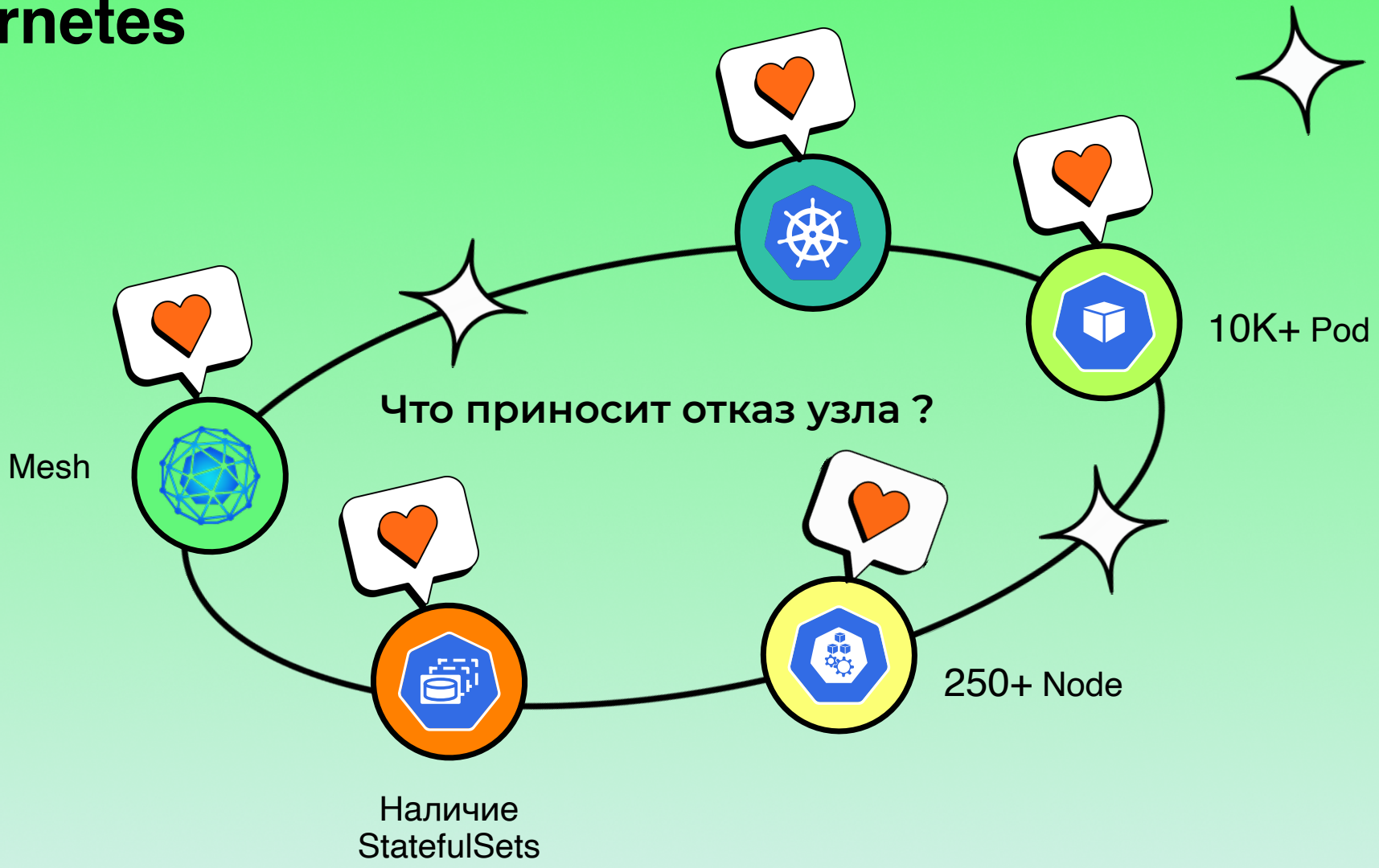


Глоссарий

Отказ узла

Событие, приводящее
к невозможности производить
запуск новых и использование
текущих Pods

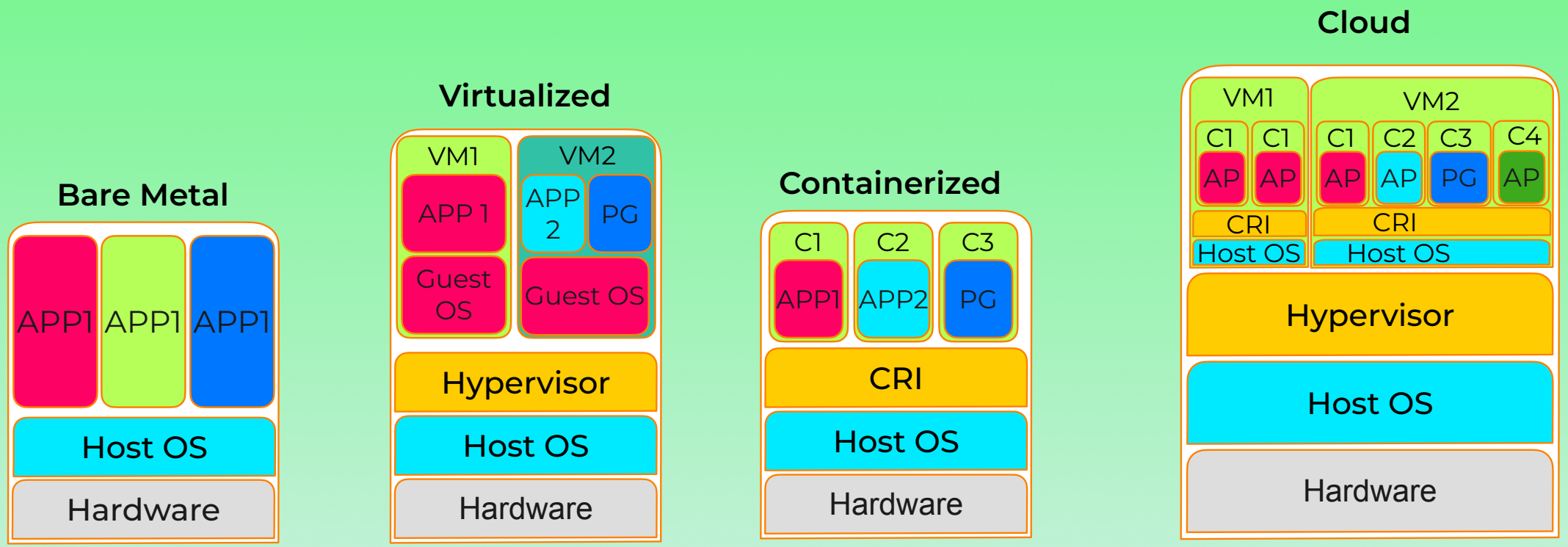
Kubernetes


10K+ Pod
Что приносит отказ узла ?
Mesh
250+ Node
Наличие
StatefulSets

Уровень абстракции в cloud

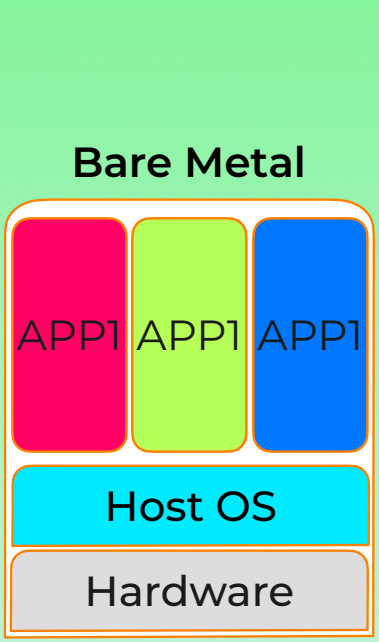
Помоги Даше найти нужный слой абстракции

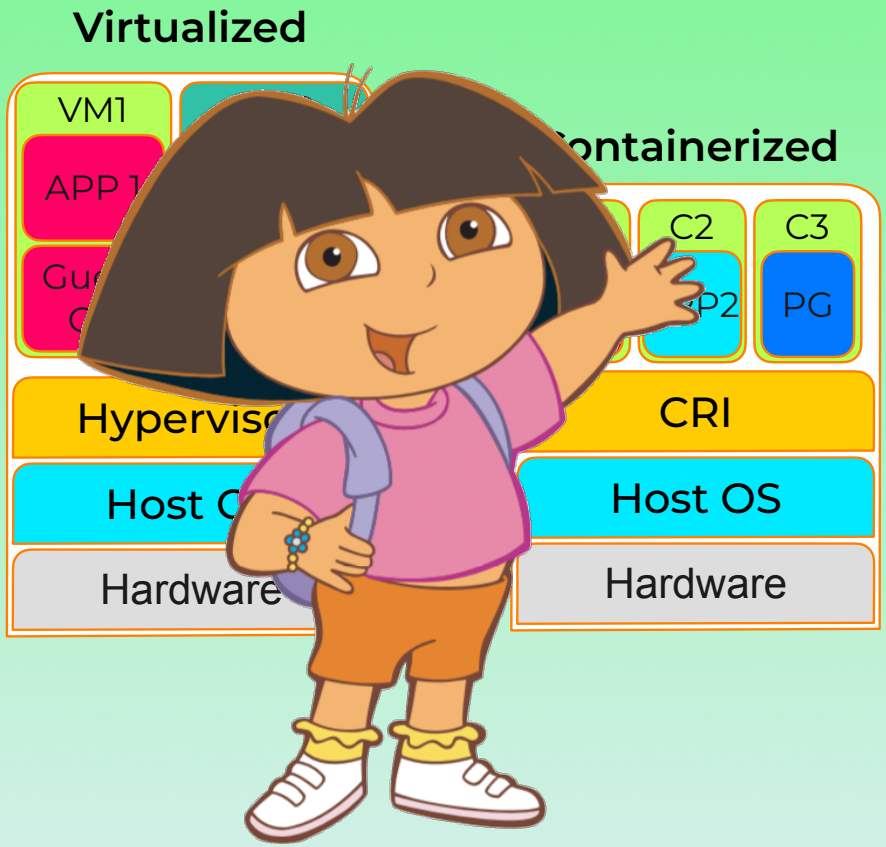
Уровень абстракции в cloud

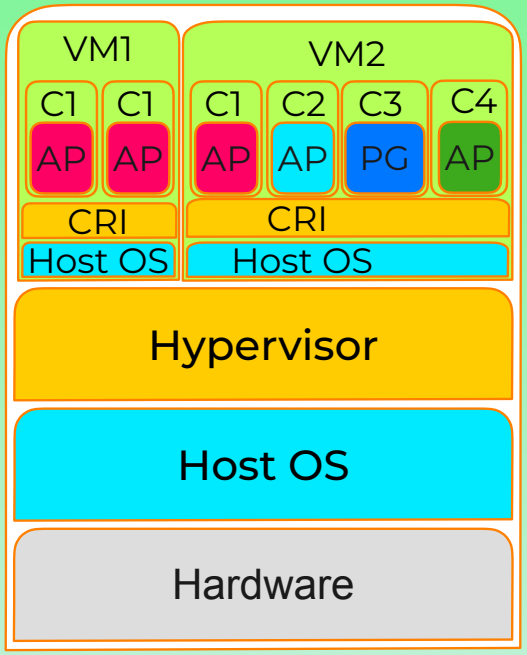


Уровень абстракции в cloud

Помоги Даше найти нужный слой абстракции







КУПЕР

Классификация отказов узла

Анализ опыта

Анализ аварий за 2 года

Классификация отказов узла

Сбой/недоступность узла

1. **Отключение узла со стороны Клауда** — 70%
   - Отказ оборудования ✖
   - Плановые работы ✖
2. **Фриз vm** — 12%
   - Миграция VM !
3. **Проблемы с Memory** — 12%
   - MemoryPresure ✖
   - System OOMKilling !
4. **Сетевая недоступность** — 3%
   - Изоляция ноды !
5. **Проблемы с FS** — 2%
   - DiskPressure ✖
   - **FS** read-only !
   - Corrupted file system !
6. **Прочие проблемы** — 1%
   - PIDProblem !
   - ProcessZ !

✖ Узел сменит статус в not **Ready**

! Узел остается в Ready

Классификация отказов узла

Сбой/недоступность узла

1. **Отключение узла со стороны Клауда** — 70%
   - Отказ оборудования
   - Плановые работы
2. **Фриз vm** — 12%
   - Миграция VM
3. **Проблемы с Memory** — 12%
   - MemoryPresure
   - System OOMKilling
4. **Сетевая недоступность** — 3%
   - Изоляция ноды
5. **Проблемы с FS** — 2%
   - DiskPressure
   - **FS** read-only
   - Corrupted file system
6. **Прочие проблемы** — 1%
   - PIDProblem
   - ProcessZ

Отказ/недоступность узла

Проблемы на узле

КУПЕР

Методы выявления отказов узла

Методы выявления отказа узла

1. Возможности из коробки

Методы выявления отказа узла

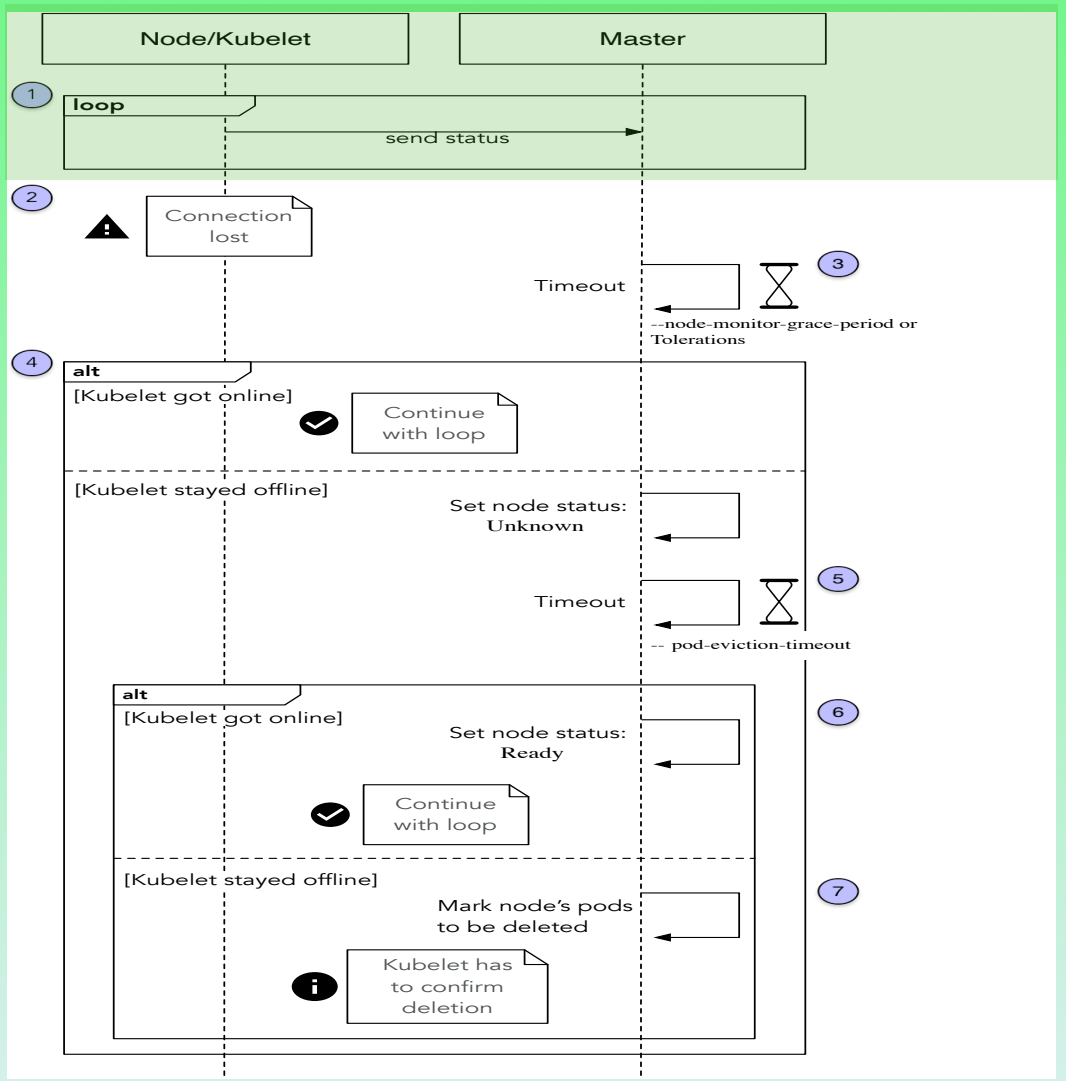
1. Возможности из коробки

kubelet

--node-status-update-frequency default 10s

c-m

-node-monitor-period default 5s



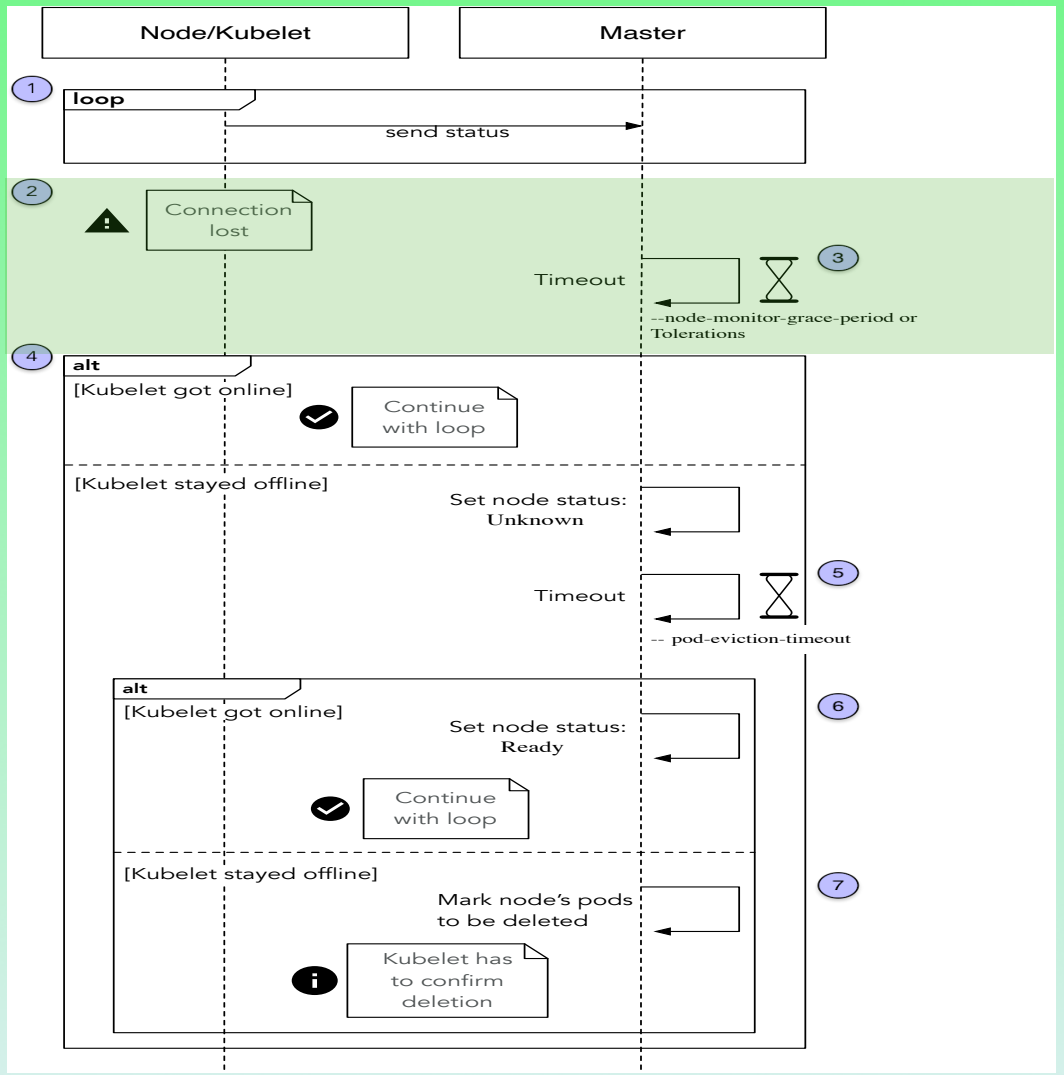
Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки

-node-status-update-frequency default 10
-nodeStatusUpdateRetry default 5

node-monitor-grace-period



Отказ/недоступность физ узла

Методы выявления отказа узла

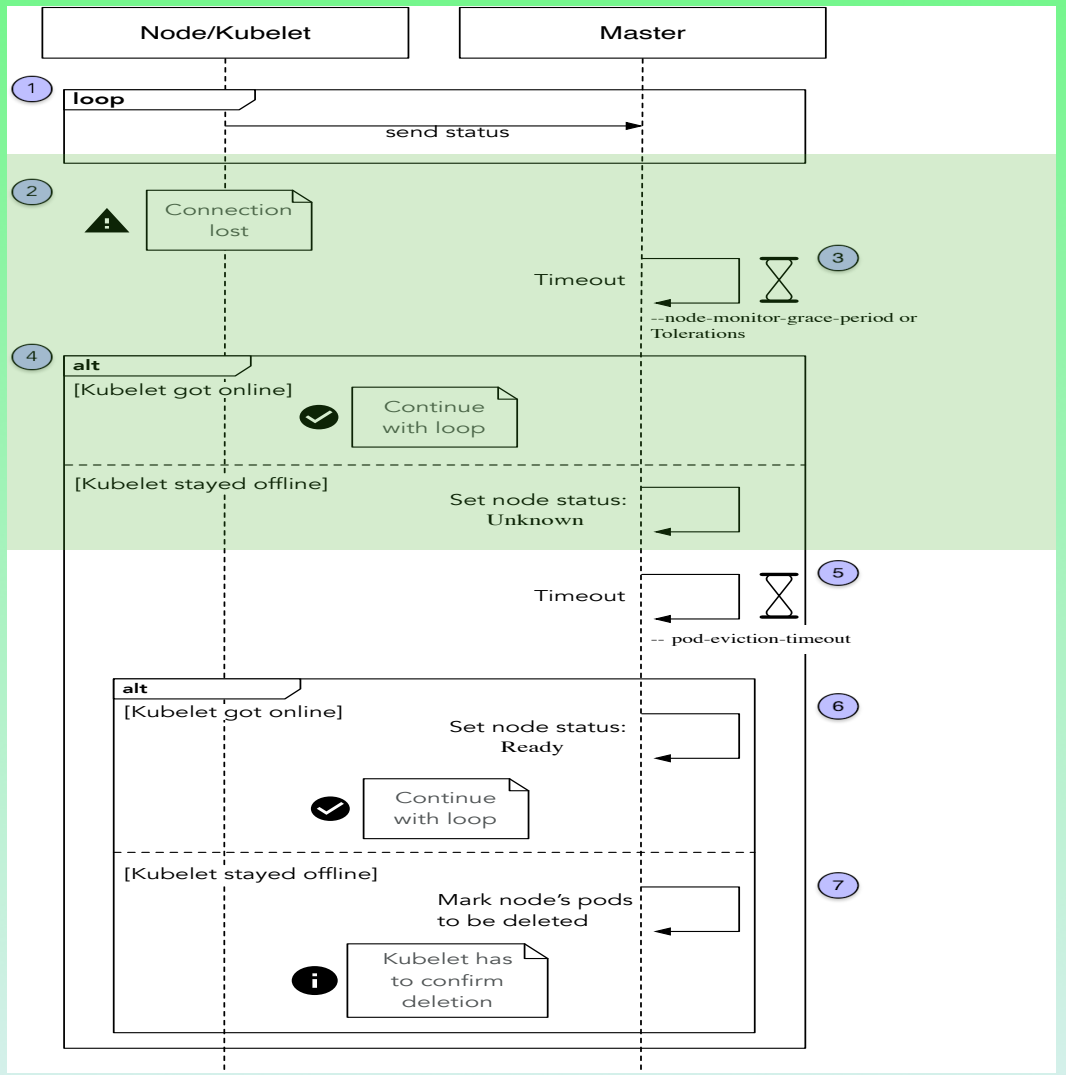
1. Возможности из коробки

Прошло 40с

Taint: node.kubernetes.io/unreachable

NodeCondition: Unknow

Проблема:
Endopoints продолжают быть доступны на протяжении 40с

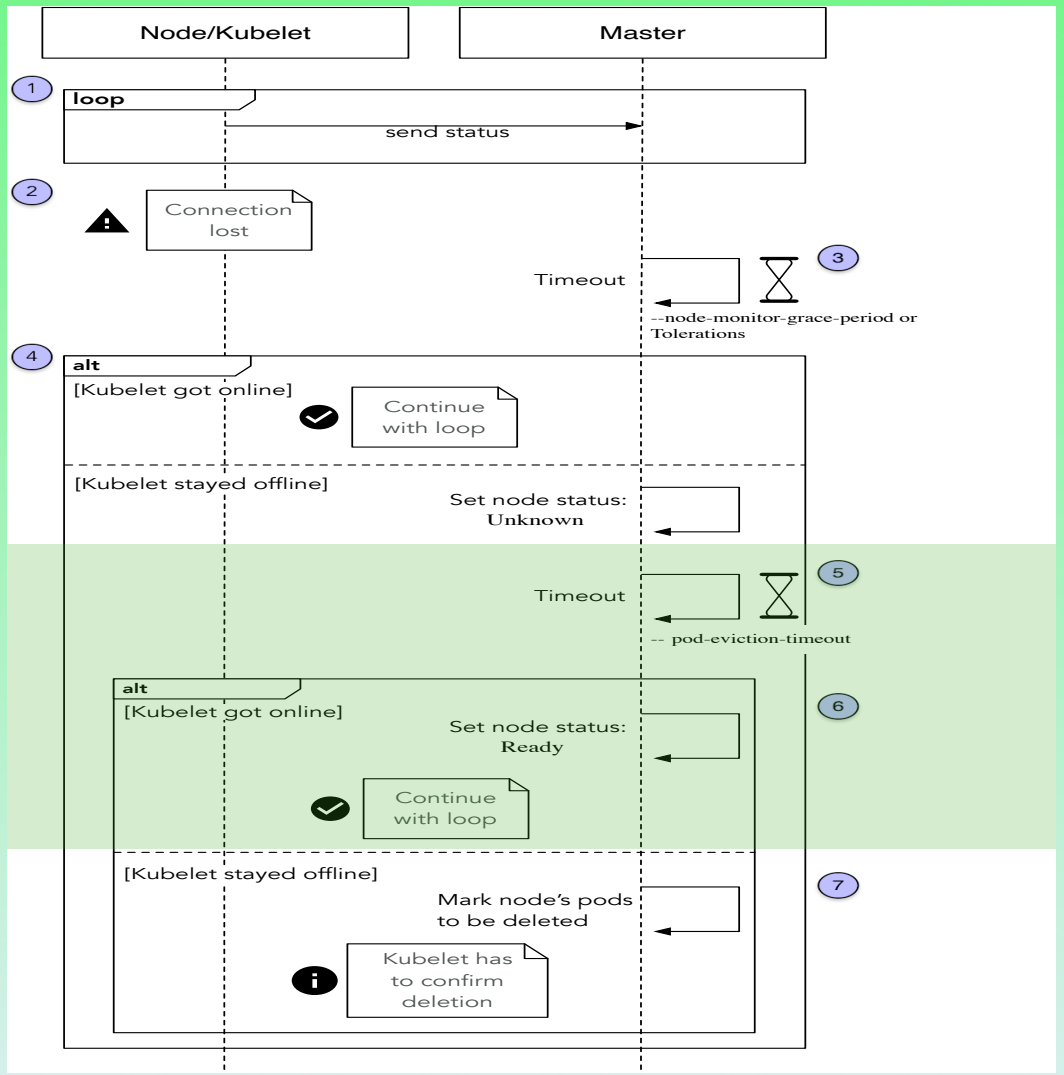


Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки

--pod-eviction-timeout default 5m0s



Отказ/недоступность физ узла

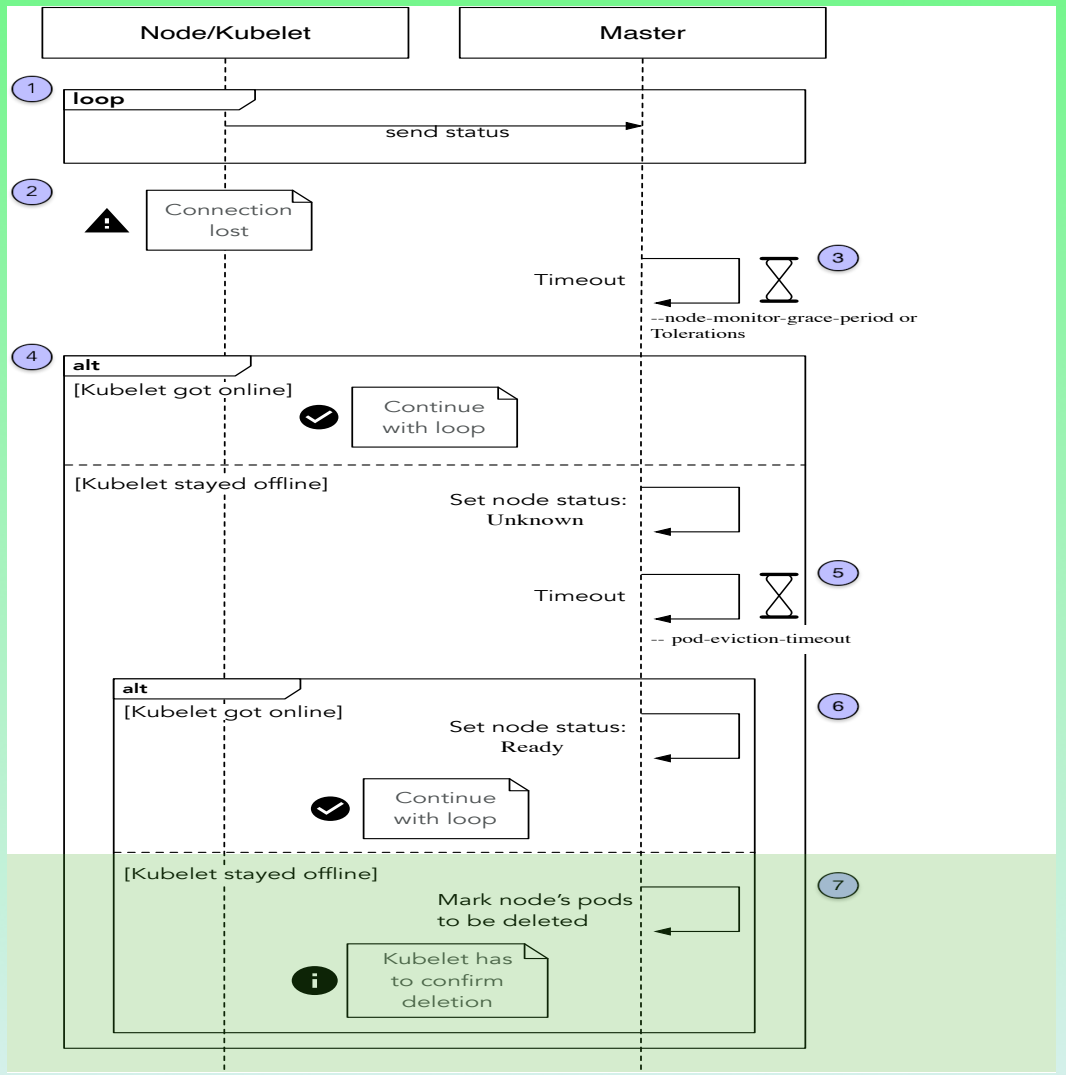
Методы выявления отказа узла

1. Возможности из коробки

Прошло 5m

запуск Eviction Pods

Проблема:
STS так и **останутся** на отказавшем узле



Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки/методы влияния

Подходит для небольших **Self Hosted** инсталляций

Преимущества Selfhosted от менеджмент решений, то что вы можете сами влиять на настройки

Уменьшаем время с 5м 40с до **1m**

node-monitor-grace-period = 16
node-status-update-frequency (4)
×
(nodeStatusUpdateRetry-1) (4)

—node-status-update-frequency=4s

—node-monitor-grace-period=16s
—node-monitor-period=2s
—pod-eviction-timeout=30s

Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки/методы влияния

Подходит для небольших **Self Hosted** инсталляций

Преимущества Selfhosted от менеджмент решений, то что вы можете сами влиять на настройки

—node-status-update-frequency=4s

—node-monitor-grace-period=16s
—node-monitor-period=2s
—pod-eviction-timeout=30s

Уменьшаем время с 5м 40с до **1m**

Но есть минус -

Доп нагрузка на API

Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки/методы влияния

Подходит для ВСЕХ решений

Можем влиять на
время eviction через Deployment

Уменьшаем время с 5м 40с до **1m**


```yaml
tolerations:
- key: node.kubernetes.io/not-ready
  effect: NoExecute
  tolerationSeconds: 20
- key: node.kubernetes.io/unreachable
  effect: NoExecute
  tolerationSeconds: 20
```

tolerationSeconds как долго этот Pod будет оставаться привязанным к узлу

Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки/методы влияния


```yaml
tolerations:
- key: node.kubernetes.io/not-ready
  effect: NoExecute
  tolerationSeconds: 20
- key: node.kubernetes.io/unreachable
  effect: NoExecute
  tolerationSeconds: 20
```

tolerationSeconds как долго этот Pod будет оставаться привязанным к узлу


```yaml
tolerations:
- key: "node.kubernetes.io/unreachable"
  operator: "Exists"
  effect: "NoExecute"
  tolerationSeconds: 6000
```

Пример: как увеличить время eviction

Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки/методы влияния

- `node.kubernetes.io/not-ready`: Узел не готов. Это соответствует NodeCondition, `Ready` равному " `False` ".
- `node.kubernetes.io/unreachable`: Узел недоступен из контроллера узла. Это соответствует NodeCondition, `Ready` равному " `Unknown` ".
- `node.kubernetes.io/memory-pressure`: Узел испытывает нехватку памяти.
- `node.kubernetes.io/disk-pressure`: Узел испытывает нехватку дискового пространства.
- `node.kubernetes.io/pid-pressure`: Узел имеет не хватку PID.
- `node.kubernetes.io/network-unavailable`: Сеть узла недоступна.
- `node.kubernetes.io/unschedulable`: Узел не подлежит планированию.

https://kubernetes.io/docs/concepts/scheduling-eviction/taint-and-toleration/

Отказ/недоступность физ узла

Методы выявления отказа узла

1. Возможности из коробки/методы влияния

Taint: **node.kubernetes.io/out-of-service**, автоматически удаляются pod с pv, на отказавших узлах.

OutOfServiceTaint поддерживается только в кластерах с k8s версии 1.26+

https://kubernetes.io/blog/2023/08/16/kubernetes-1-28-non-graceful-node-shutdown-ga/

Отказ/недоступность физ узла

Методы выявления отказа узла

2. Node Problem Detector

https://github.com/kubernetes/node-problem-detector

Node-Problem-Detector (NPD)

- помогает обнаруживать проблемы на узлах
- сообщает о проблемах с узлом серверу api.
- отдает метрики формата Prometheus

NPD входит в поставку большинства Клауд менеджемент Kubernetes

Методы выявления отказа узла

2. Node Problem Detector / Принцип действия

Проверка компонентов и системных метрик

Выявление проблем из логов компонентов

Мониторинг метрик через /proc/*

Реагирование на events

Прослушивание /dev/kmsg и выявление событий по вхождению regex

Запуск скриптов/приложений

Запуск предустановленных скриптов для анализа

Благодаря чему можно получать events от сервиса метаинфрмации облака

Методы выявления отказа узла

2. Node Problem Detector / Функционал

Функции проверки		
DiskReadonly	FrequentContainerdRestart	PIDPressure
SpotPriceNodeReclaimNotification	FrequentKubeletRestart	CNIProblem
ScheduledEvent	KUBELETProblem	PIDProblem
MemoryProblem	NTPProblem	FDProblem
ProcessD	KUBEPROXYProblem	EmptyDirVolumeGroupStatusError
MountPointProblem	ProcessZ	DiskProblem
ConntrackFullProblem	LocalPvVolumeGroupStatusError	MemoryPressure
	CRIProblem	DiskPressure

Методы выявления отказа узла

2. Node Problem Detector / Функционал / запланированные работы Cloud

HUAWEI CLOUD

Yandex Cloud

MetaData

Публикация о предстоящих событиях на уровне Compute

- FreezeScheduled
- RebootScheduled
- TerminateScheduled
- PreemptScheduled

Методы выявления отказа узла

2. Node Problem Detector / Функционал / запланированные работы Cloud

Feature	IMDS Processor	Queue Processor
Spot Instance Termination Notifications (ITN)	✅	✅
Scheduled Events	✅	✅
Instance Rebalance Recommendation	✅	✅
ASG Termination Lifecycle Hooks	✅	✅
AZ Rebalance Recommendation	❌	✅
Instance State Change Events	❌	✅

https://github.com/aws/aws-node-termination-handler

Методы выявления отказа узла

2. Node Problem Detector / Функционал / запланированные работы Cloud

Yandex Cloud

Feature
Spot Instance Termination Notifications (ITN)
Scheduled Events
Instance State Change Events

https://yandex.cloud/ru/docs/compute/concepts/vm-policies

Методы выявления отказа узла

3. Операторы

Методы выявления отказа узла

3. Операторы / medik8s

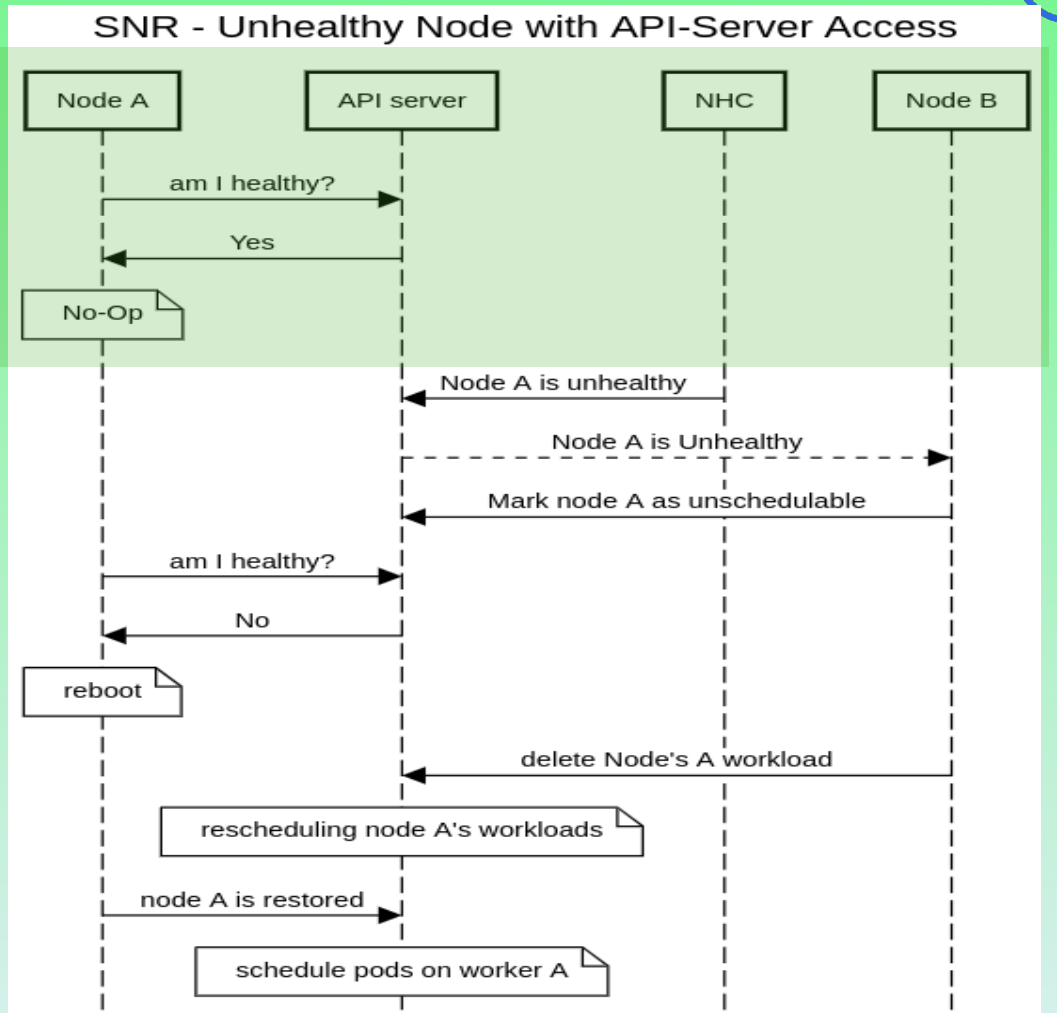
Self Node Remediation

Fence Agents Remediation

Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

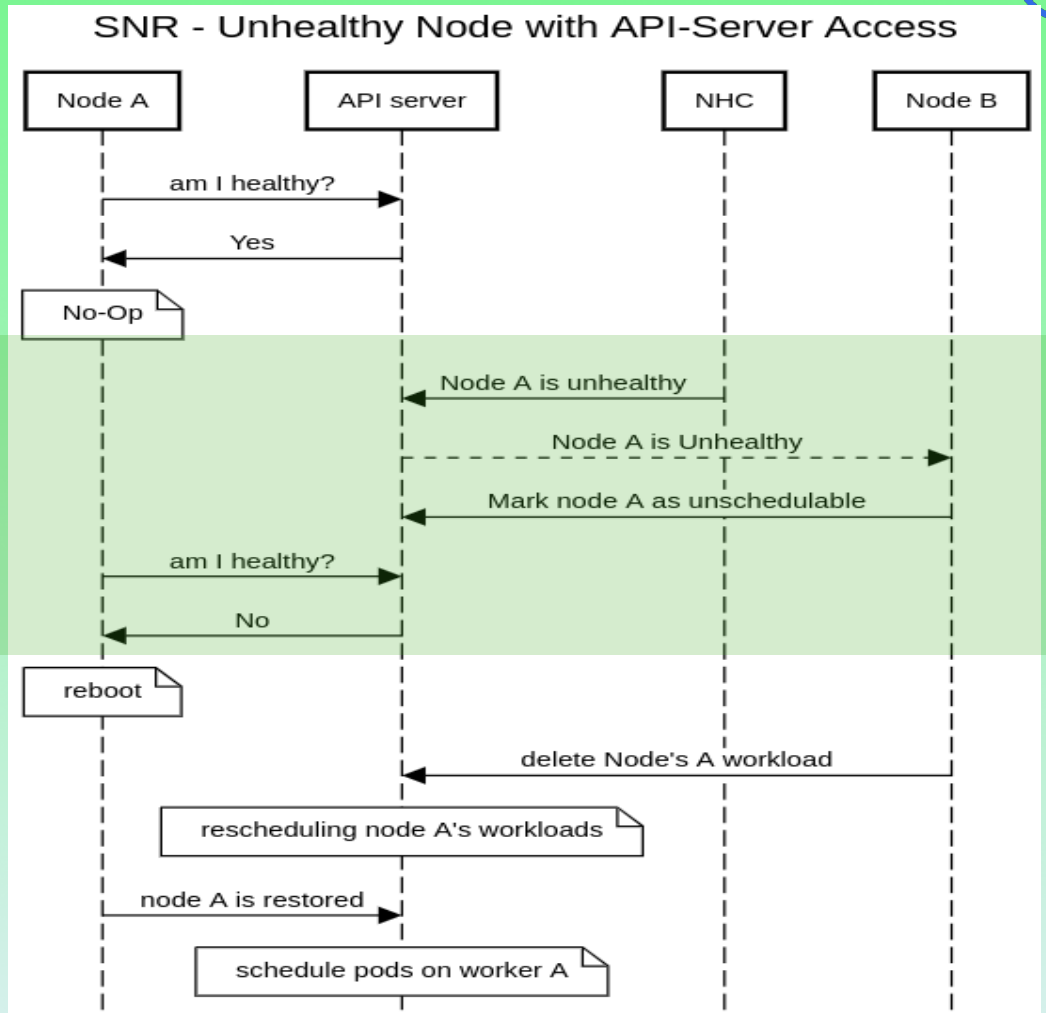
Узел с доступом к API-серверу



Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

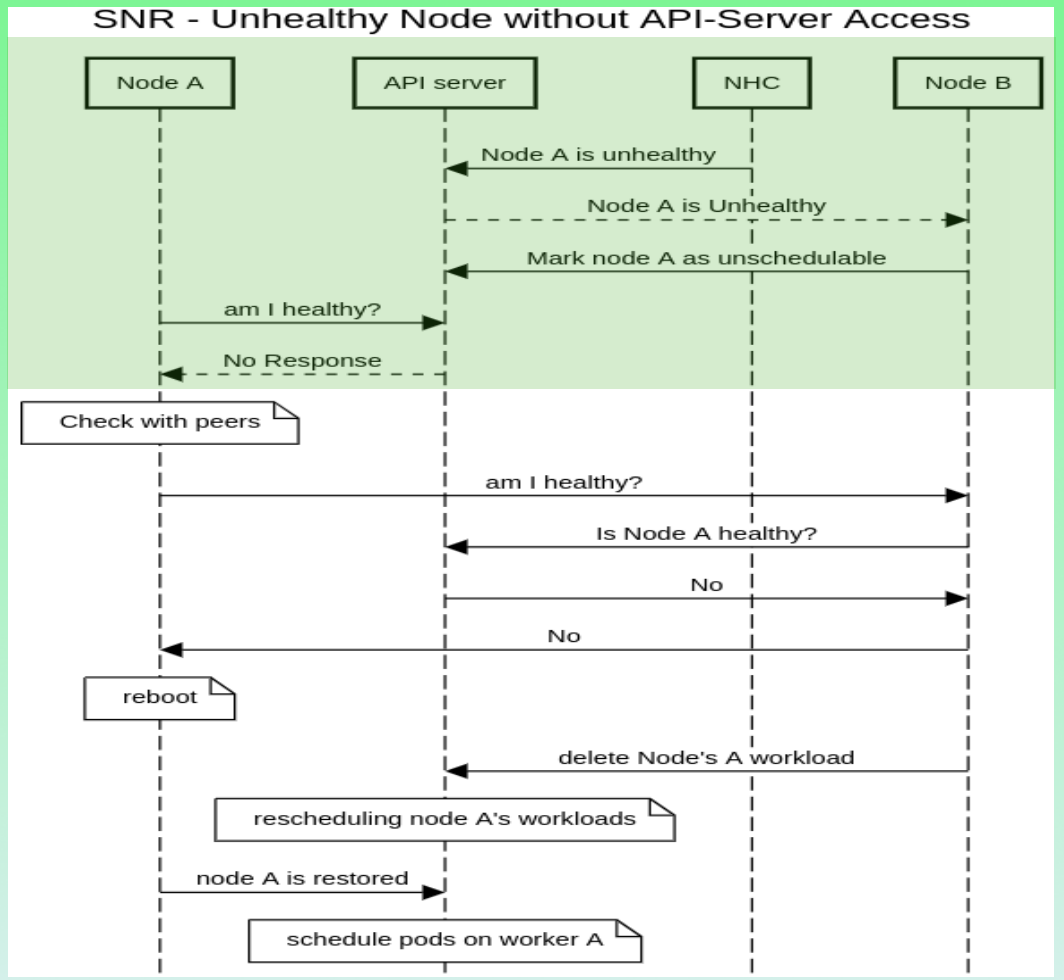
Узел с доступом к API-серверу



Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

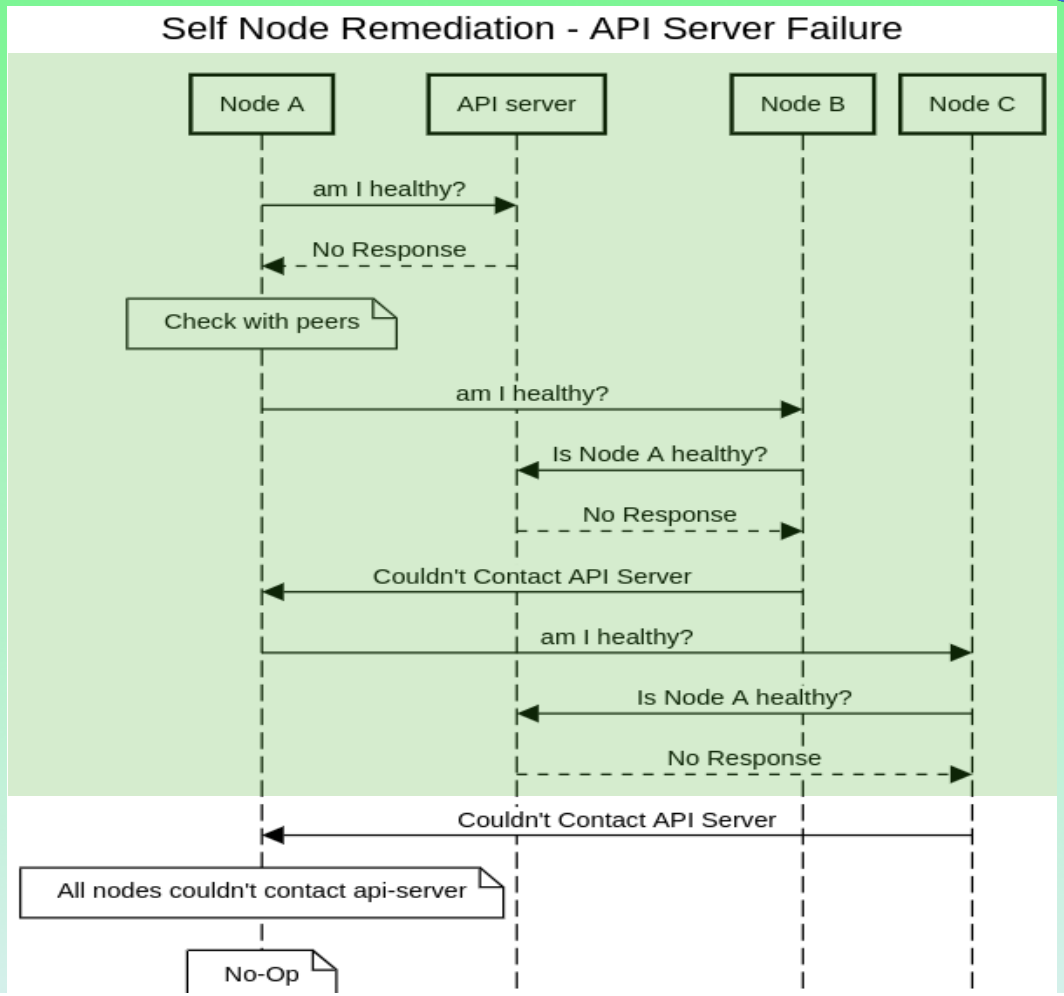
Узел без доступа к API-серверу



Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

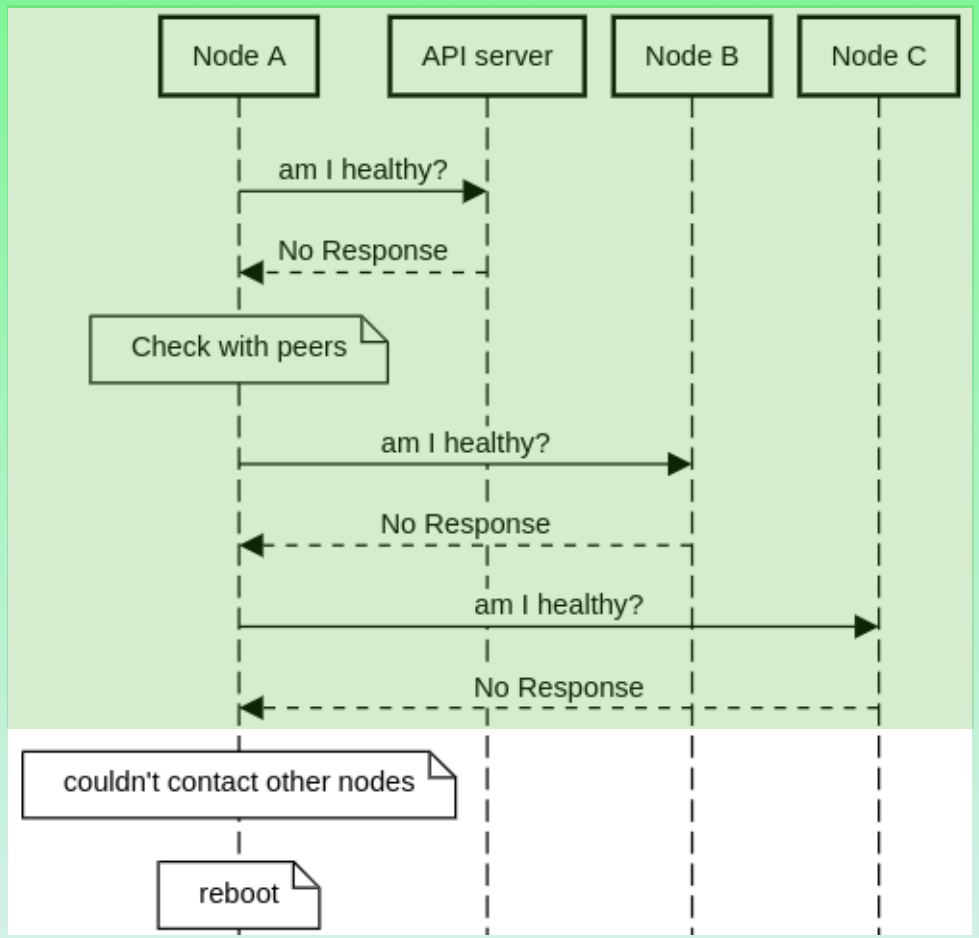
Узел без доступа к API-серверу



Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

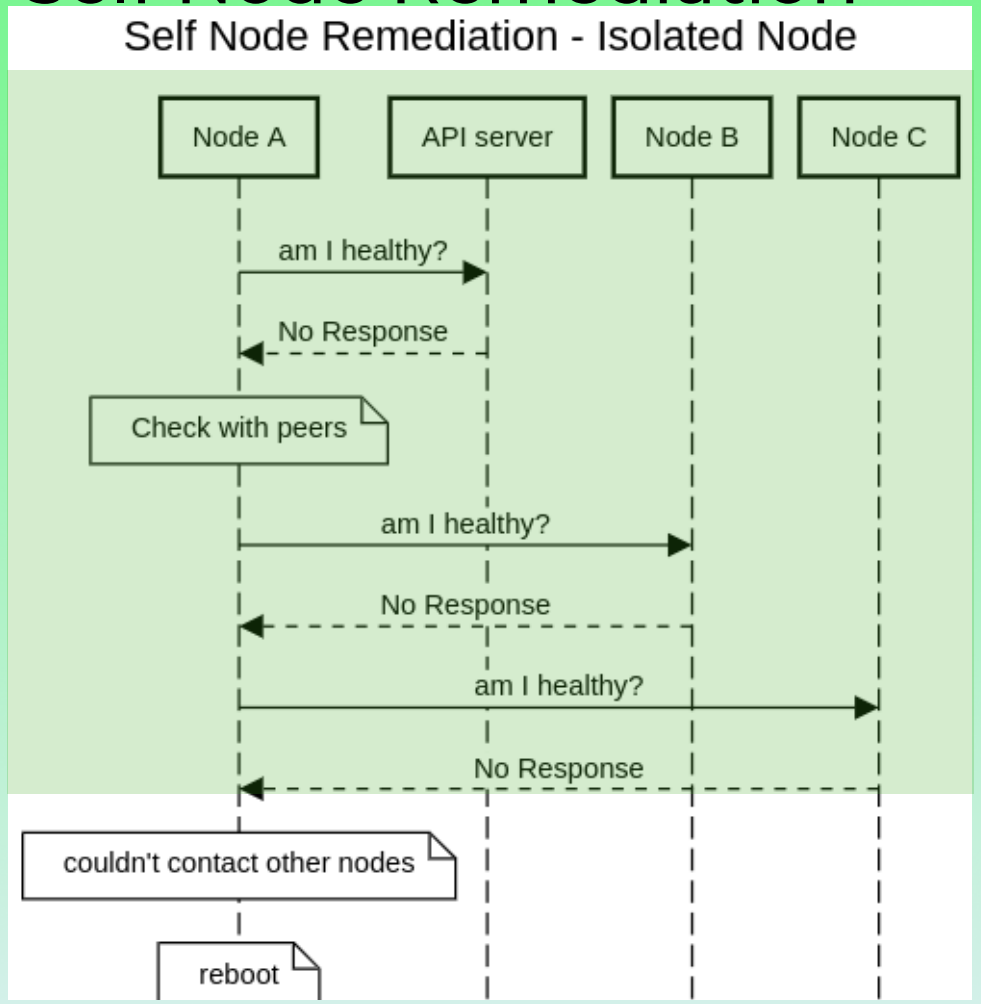
Изолированный узел



Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

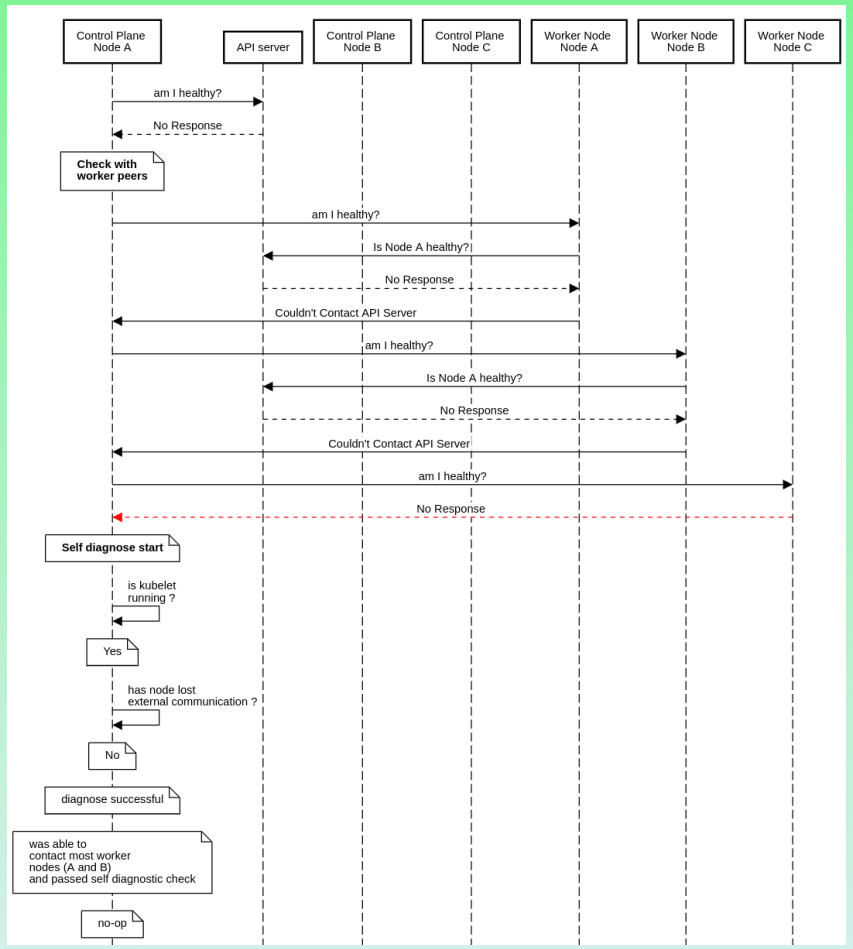
Изолированный узел



Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

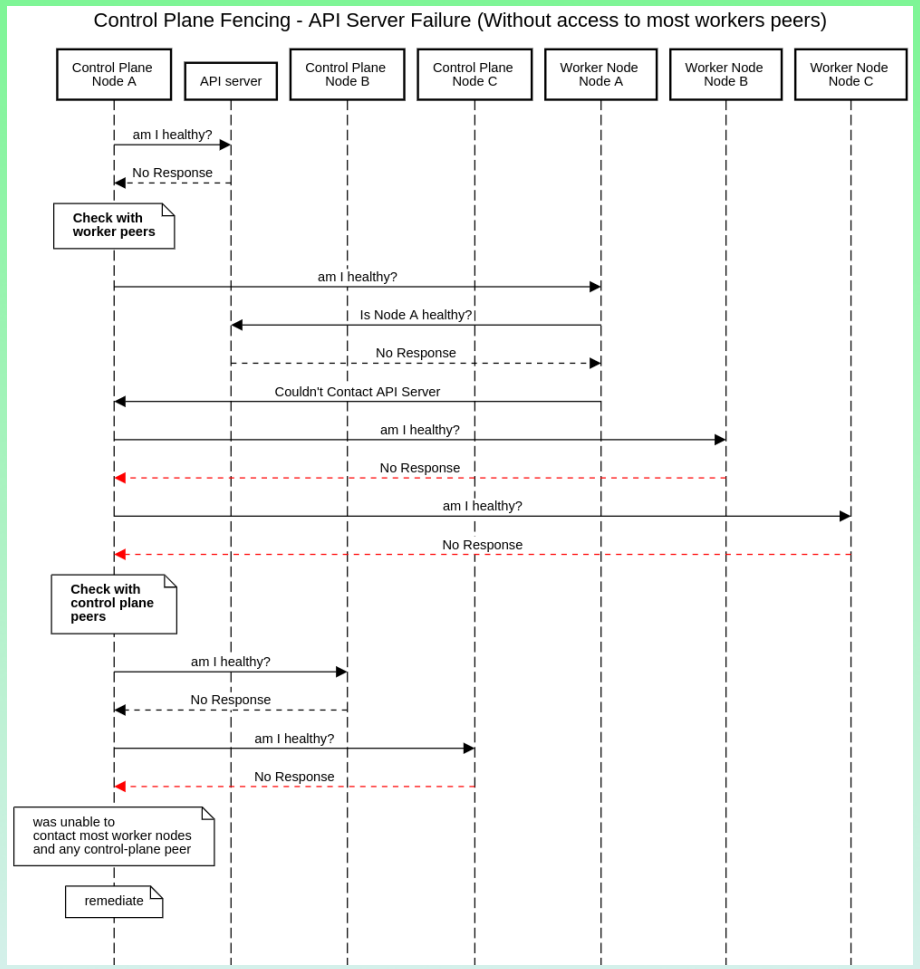
Узел ControlPlane / доступ к узлам есть



Методы выявления отказа узла

3. Операторы / medik8s / Self Node Remediation

Узел ControlPlane /
доступ к узлам отсутствует



Методы выявления отказа узла

3. Операторы / medik8s / Fence Agents Remediation

Работа с узлом

1

- Добавляет Taint
- Запуск eviction

Ограждает fence agent

2

ClusterLabs / fence-agents

https://github.com/ClusterLabs/fence-agents

Рестарт & Удаление узела

3

Методы выявления отказа узла

3. Операторы / kube-fencing

https://github.com/ClusterLabs/fence-agents

Методы выявления отказа узла

3. Операторы / kube-fencing

https://github.com/ClusterLabs/fence-agents

fencing-controller

1

- Следит за статусом узлов

fencing-switcher

2

- Работает на узле

fencing-agents

3

- Работа с агентами

Методы выявления отказа узла

3. Операторы / kube-fencing

https://github.com/ClusterLabs/fence-agents

fencing-controller

1. Следит за статусом узлов

fencing-switcher

2. Работает на узле

fencing-agents

3. Работа с агентами

Fensing-agens							
fence_ack_manual	fence_brocade	fence_dummy	fence_idrac	fence_ilo4_ssh	fence_ipmilan	fence_ovh	fence_rsb
fence_alom	fence_cisco_mds	fence_eaton_snmp	fence_ifmib	fence_ilo_moonshot	fence_ironic	fence_powerman	fence_sanbox2
fence_amt	fence_cisco_ucs	fence_emerson	fence_ilo	fence_ilo_mp	fence_kdump	fence_pve	fence_sbd
fence_apc	fence_compute	fence_eps	fence_ilo2	fence_ilo_ssh	fence_ldom	fence_raritan	fence_scsi
fence_apc_snmp	fence_docker	fence_hds_cb	fence_ilo3	fence_imm	fence_lpar	fence_rcd_serial	fence_tripplite_snmp
fence_azure_arm	fence_drac	fence_hpblade	fence_ilo3_ssh	fence_intelmodular	fence_mpath	fence_rhevm	fence_vbox
fence_bladecenter	fence_drac5	fence_ibmblade	fence_ilo4	fence_ipdu	fence_netio	fence_rsa	fence_virsh
			fence_wti	fence_vmware_soap	fence_vmware	fence_xenapi	fence_zvmip

Методы выявления отказа узла

3. Операторы / medik8s / Node Maintenance Operator

https://github.com/medik8s/node-maintenance-operator

1

```yaml
apiVersion: nodemaintenance.medik8s.io/v1beta1
kind: NodeMaintenance
metadata:
  name: nodemaintenance-sample
spec:
  nodeName: node02
  reason: "Test node maintenance"
```

2

Запуск Drain

- msg="Maintenance taints will be added to node node02”
- msg="Applying medik8s.io/drain taint add on Node: node02”
- msg="Patching taints on Node: node02"

Методы выявления отказа узла

3. Операторы / medik8s / draino

https://github.com/planetlabs/draino

1. Следит за статусом узлов
2. Блокирует планирование Cordon
3. Запускает Drain

Методы выявления отказа узла

4. Решения на базе метрик

Prometheus

+

Alertmanager

+

+ **Draino**

Методы выявления отказа узла

4. Решения на базе метрик

https://github.com/cloudflare/sciuro

1

```
alert: NodeUpTooLong
expr: (time() - node_boot_time_seconds) / 60 / 60 / 24 > 7
labels:
  notify: node-condition-k8s
  priority: "8"
annotations:
  description: Node '{{ $labels.instance }}' has been up for more than 7 days
  summary: Node '{{ $labels.instance }}' uptime too long
```

2

```
$ kubectl get node worker01 -o json | jq '.status.conditions[] | select(.type | test("^AlertManager_"))'
{
  "lastHeartbeatTime": "2024-10-01T16:07:10Z",
  "lastTransitionTime": "2024-10-01T15:34:07Z",
  "message": "[P8] Node 'worker01' uptime too long",
  "reason": "AlertIsFiring",
  "status": "True",
  "type": "AlertManager_NodeUpTooLong"
}
```

Сравнение подходов

Встроенные возможности

Node Problem Detector

Операторы

Решения на базе метрик

Сравнение подходов

Встроенные возможности

✓Работает из коробки
✓Возможность подстроки
✓Фиксация изоляции/недостпуности узла
✓Поддержка node Conditions
✓Отдает метрики

- Отсутствует фиксация проблем внутри узла
- Отсутствует фиксация оповещений от cloud
- Возможен аффект API

Node Problem Detector

Операторы

Решения на базе метрик

Сравнение подходов

Встроенные возможности

Node Problem Detector

✓Фиксирует проблему внутри узла
✓Фиксирует работы cloud
✓Модульный
✓Поддержка node Conditions
✓Отдает метрики

- Работает пока работает узел
- Требуется привилегированный режим

Операторы

Решения на базе метрик

Сравнение подходов

Встроенные возможности

Node Problem Detector

Операторы

✓Фиксирует проблему внутри узла
✓Фиксирует недоступность/ изоляцию узла
✓Автоматичиский запуск Cordon/ Drain

- Отсутствует фиксация оповещений от cloud
- Требуется привилегированный режим
- Отсутствует фиксация проблем внутри узла

Решения на базе метрик

Сравнение подходов

Встроенные возможности

Node Problem Detector

Операторы

Решения на базе метрик

✓Работает с метриками
✓Прост в эксплуатации

- Требуются инфраструктура Сбора метрик
- Требуются сервисы создающие метрики
- Требуются Операторы (Draino, etc)

Сравнение подходов

- Встроенные возможности
- Node Problem Detector
- Операторы
- Решения на базе метрик

1 Отключение узла со стороны Клауда
- Отказ оборудования
- Плановые работы

2 Фриз vm
- Миграция VM

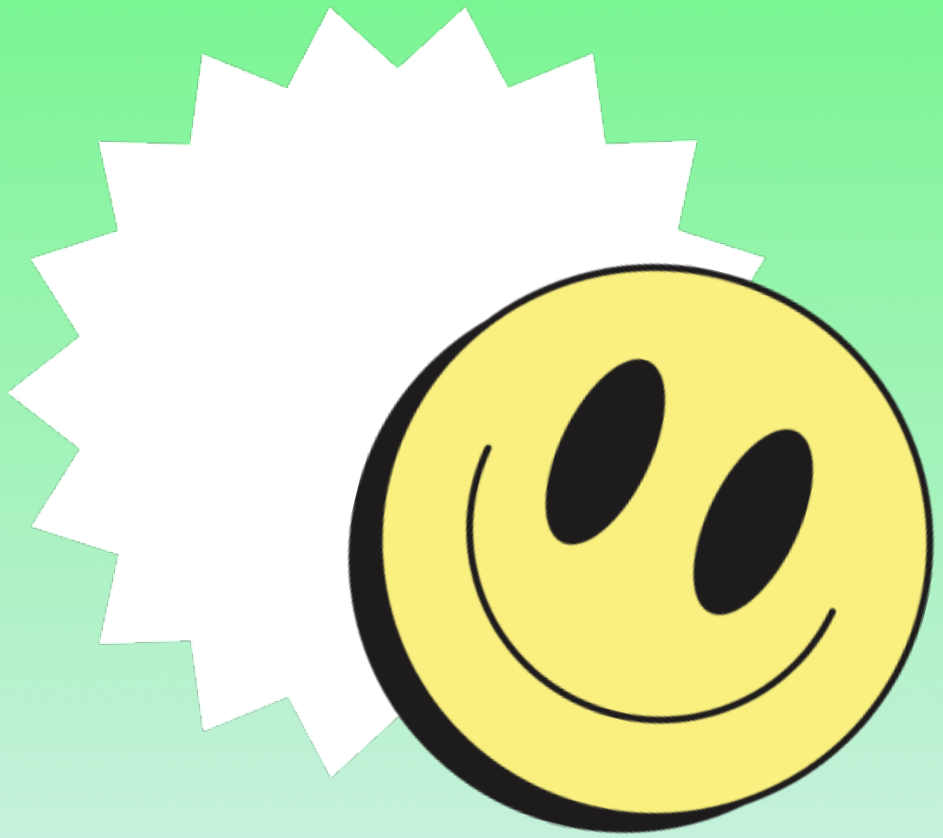
3 Проблемы с Memory
- MemoryPresure
- System OOMKilling

4 Сетевая недоступность
- Изоляция ноды

5 Проблемы с FS
- DiskPressure
- **FS** read-only
- Corrupted file system

6 Прочие проблемы
- PIDProblem
- ProcessZ

Приоритизируем вывод/запуск Pods



Приоритизируем вывод/запуск Pods

PriorityClass

- Приоритет при вытеснении (при отказе узла)
- Приоритет при планировании/запуске
- Выделение ресурсов при их нехватке (при планирование)

Чек-лист

По уменьшению влияния инфраструктуры на ваш сервис

Уровень Влияния	Задача	Используемая технология
Cloud	Распределение воркер нод из одной группы по разным стойкам и физ нодам	GPE/AWS/Azure/YCloud: placement groups VMware: affinity rules
Кластер	Распределение pod по зонам доступности	nodeAntiAffinity & topologySpreadConstraints
Сервис	Распределение pod по разным узлам	podAntiAffinity
Сервис	Гарантия доступности N pods	PodDisruptionBudget
Сервис	Ускорение запуска Pod	PriorityClasses

купер

Дмитрий Рыбалка

Техлид отдела базовой ИТ-инфраструктуры в Купере