

Ниже Kubernetes: Расследования в GPU-кластерах на Talos Linux

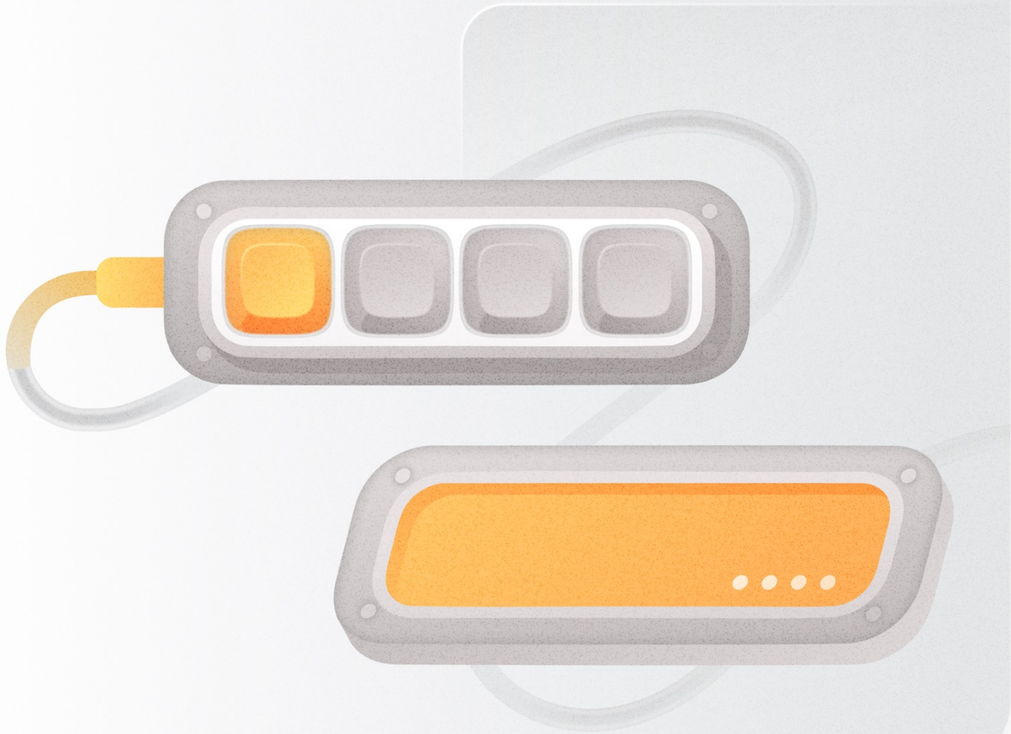
Антон Бурулев

Ведущий инженер,
Spirit Runtime Compute

- Делал для SberCloud GPU Kubernetes по кнопке для DMZ зон
- GPU в Baremetal Kubernetes впервые потрогал в серверной Ашана
- В Т-Банке строю платформу на базе Kubernetes, в том числе и для запуска GPU нагрузки



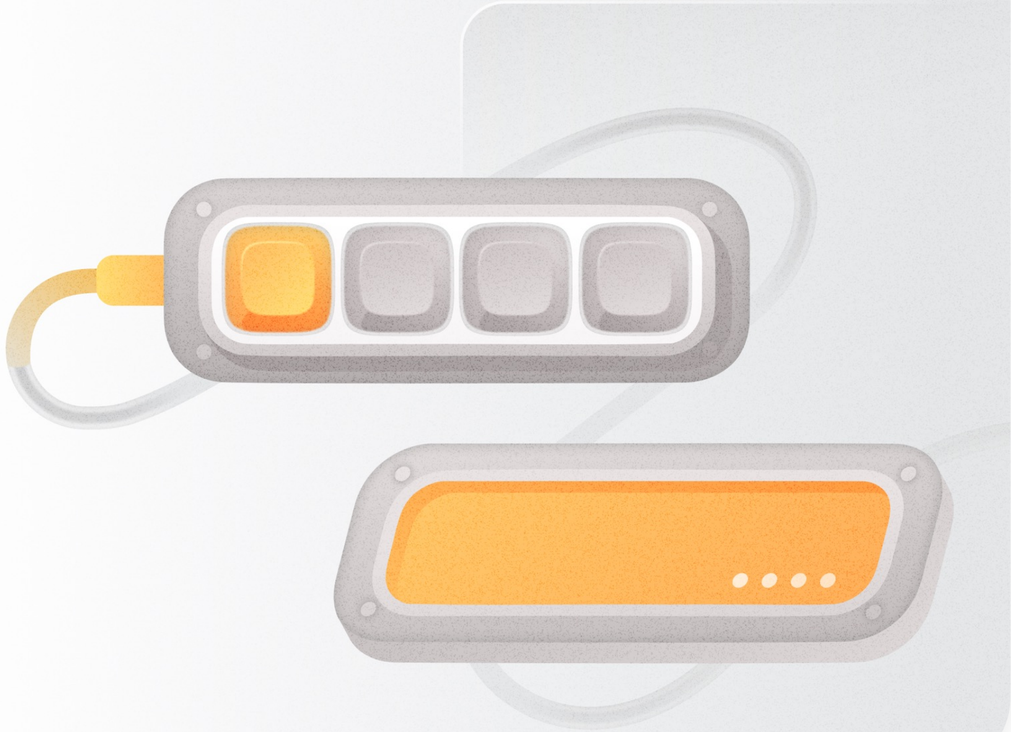
Spirit Runtime Compute



Строим K8S aaS

- Отвечаем за K8S и все что его окружает – ОС, GPU, CNI, CSI и тд
- Строим автоматизированную поставку на базе Cluster API

Spirit Runtime Compute



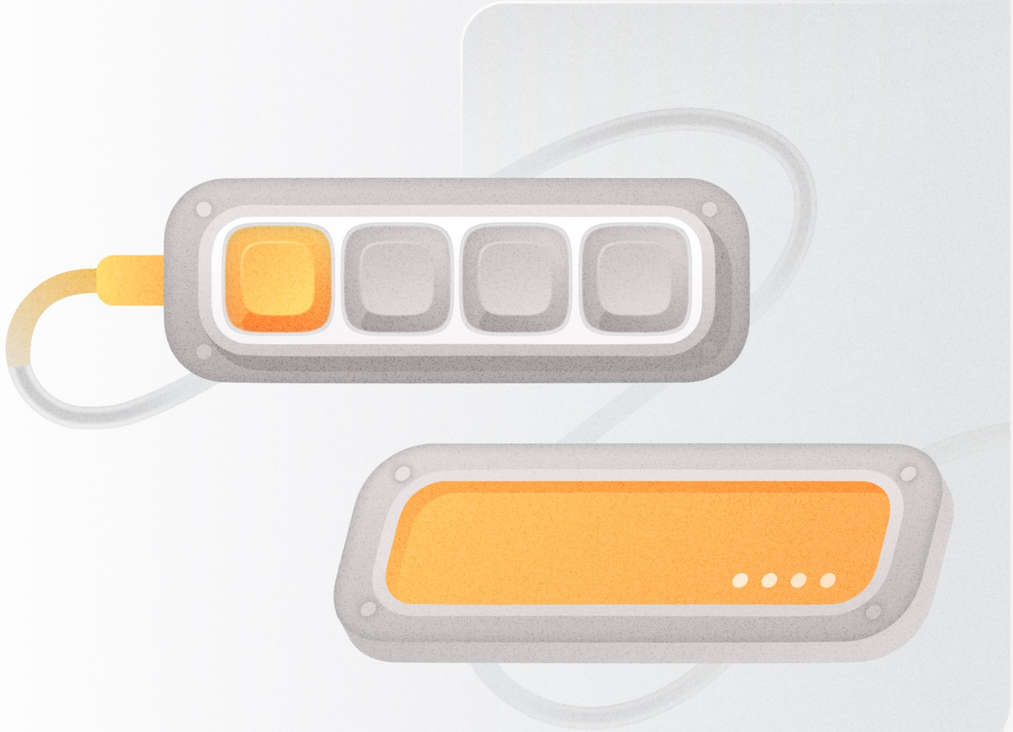
Строим K8S aaS

- Отвечаем за K8S и все что его окружает – ОС, GPU, CNI, CSI и тд
- Строим автоматизированную поставку на базе Cluster API

Гетерогенная инфраструктура

- ОС – Ubuntu, Talos
- On-prem, Cloud, Baremetal

Spirit Runtime Compute



Строим K8S aaS

- Отвечаем за K8S и все что его окружает – ОС, GPU, CNI, CSI и тд
- Строим автоматизированную поставку на базе Cluster API

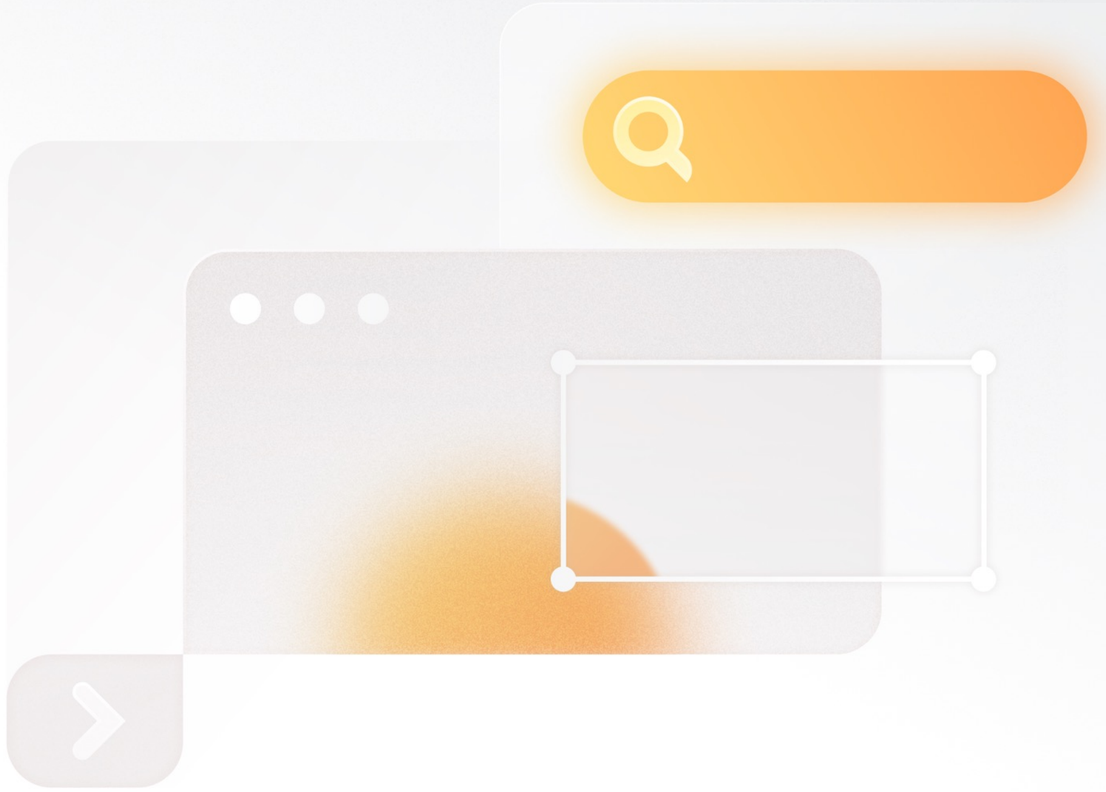
Гетерогенная инфраструктура

- ОС – Ubuntu, Talos
- On-prem, Cloud, Baremetal

Цифры

- 250k+ vCPU, 1.5 PB RAM
- 300k подов
- 3 года на Talos в Production

Содержание



01 GPU до Talos

02 GPU и Talos

03 Проблема 0 – Первая встреча с LLM

04 Проблема 1 – Медленный запуск

05 Проблема 2 – Высокий latency

06 Выводы

GPU до Talos

Виртуальные машины с Ubuntu

- 1 BM – 1 GPU
- Каждый pod может получить 1 GPU

GPU до Talos

Виртуальные машины с Ubuntu

- 1 BM – 1 GPU
- Каждый pod может получить 1 GPU

Простота установки

- NVIDIA GPU Operator
- Единственная сложность – сборка драйвера под ОС

NVIDIA GPU Operator - Компоненты

01 **Operator**

Управляет всем жизненным циклом

GPU в кластере

NVIDIA GPU Operator - Компоненты

01 **Operator**

Управляет всем жизненным циклом

GPU в кластере

02 **Driver**

Интеграция GPU с ОС

NVIDIA GPU Operator - Компоненты

01 **Operator**

Управляет всем жизненным циклом GPU в кластере

02 **Driver**

Интеграция GPU с ОС

03 **Container Toolkit**

Интеграция GPU с Container Runtime

NVIDIA GPU Operator - Компоненты

01 **Operator**

Управляет всем жизненным циклом GPU в кластере

02 **Driver**

Интеграция GPU с ОС

03 **Container Toolkit**

Интеграция GPU с Container Runtime

04 **Device Plugin**

Интеграция GPU устройств с kubelet

NVIDIA GPU Operator - Компоненты

01 **Operator**

Управляет всем жизненным циклом GPU в кластере

02 **Driver**

Интеграция GPU с ОС

03 **Container Toolkit**

Интеграция GPU с Container Runtime

04 **Device Plugin**

Интеграция GPU устройств с kubelet

05 **Monitoring & Discovery**

Инвентаризация и сбор метрик с GPU

NVIDIA GPU Operator - Компоненты

01 **Operator**

Управляет всем жизненным циклом GPU в кластере

02 **Driver**

Интеграция GPU с ОС

03 **Container Toolkit**

Интеграция GPU с Container Runtime

04 **Device Plugin**

Интеграция GPU устройств с kubelet

05 **Monitoring & Discovery**

Инвентаризация и сбор метрик с GPU

06 **Дополнительные компоненты**

vGPU, MIG, MPS и т.д.

Переход в Baremetal Talos

00

Сохранение текущих способов поставки

Sidero – готовый инфраструктурный провайдер для Cluster API

Переход в Baremetal Talos

00

Сохранение текущих способов поставки

Sidero – готовый инфраструктурный провайдер для Cluster API

01

Иммутабельность

- Rootfs – readonly
- Нет shell, package manager

Переход в Baremetal Talos

00

Сохранение текущих способов поставки

Sidero – готовый инфраструктурный провайдер для Cluster API

01

Иммутабельность

- Rootfs – readonly
- Нет shell, package manager

02

Декларативное управление

- YAML конфигурация ОС
- Набор контроллеров как в Kubernetes

Переход в Baremetal Talos

03

Безопасность

- mTLS с короткоживущими сертификатами
- Hardening ядра

Переход в Baremetal Talos

03

Безопасность

- mTLS с короткоживущими сертификатами
- Hardening ядра

04

Kubernetes only

- Автоматизация базовых действий
- Управление etcd

Переход в Baremetal Talos

03

Безопасность

- mTLS с короткоживущими сертификатами
- Hardening ядра

04

Kubernetes only

- Автоматизация базовых действий
- Управление etcd

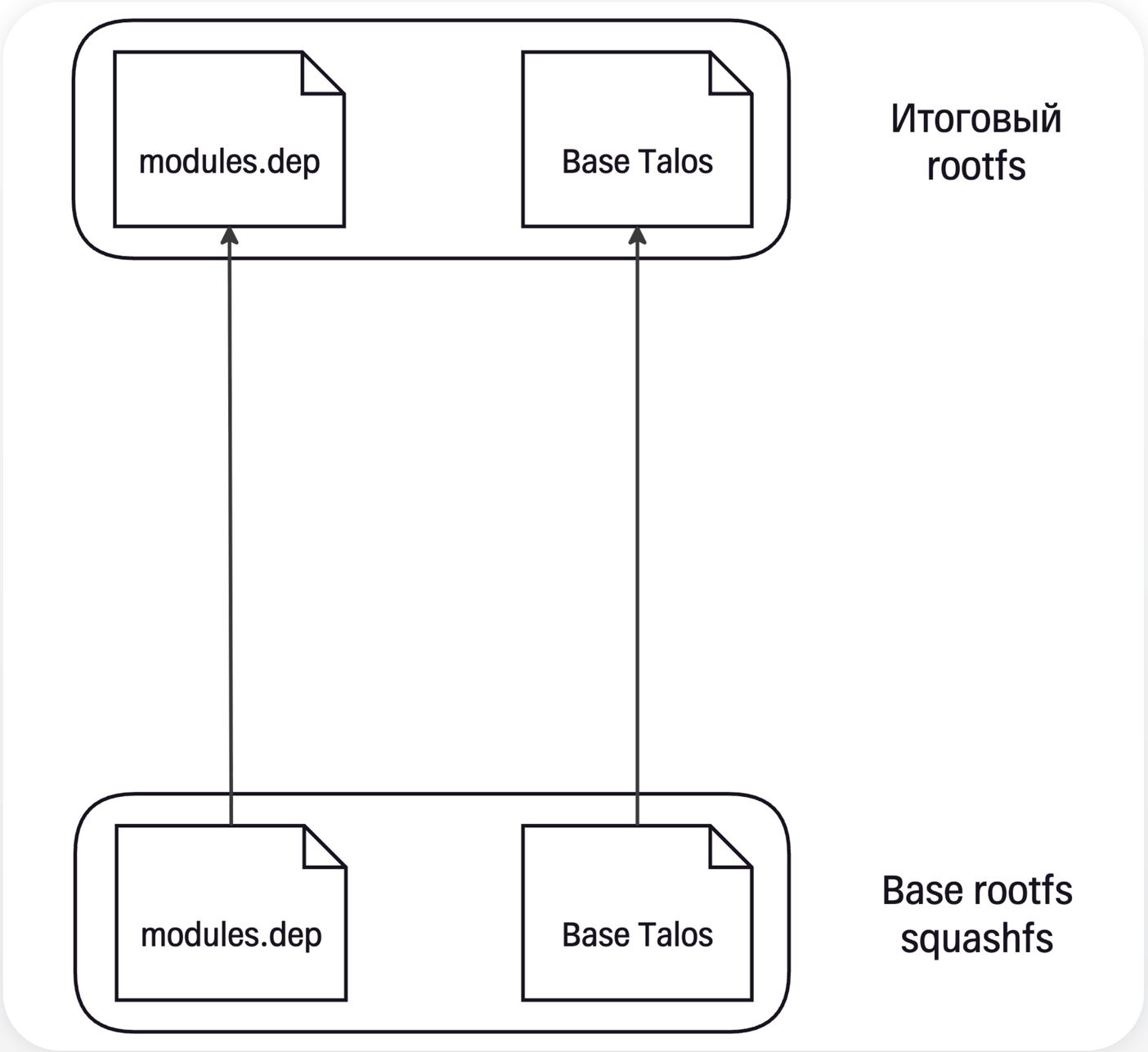
05

Эфемерность

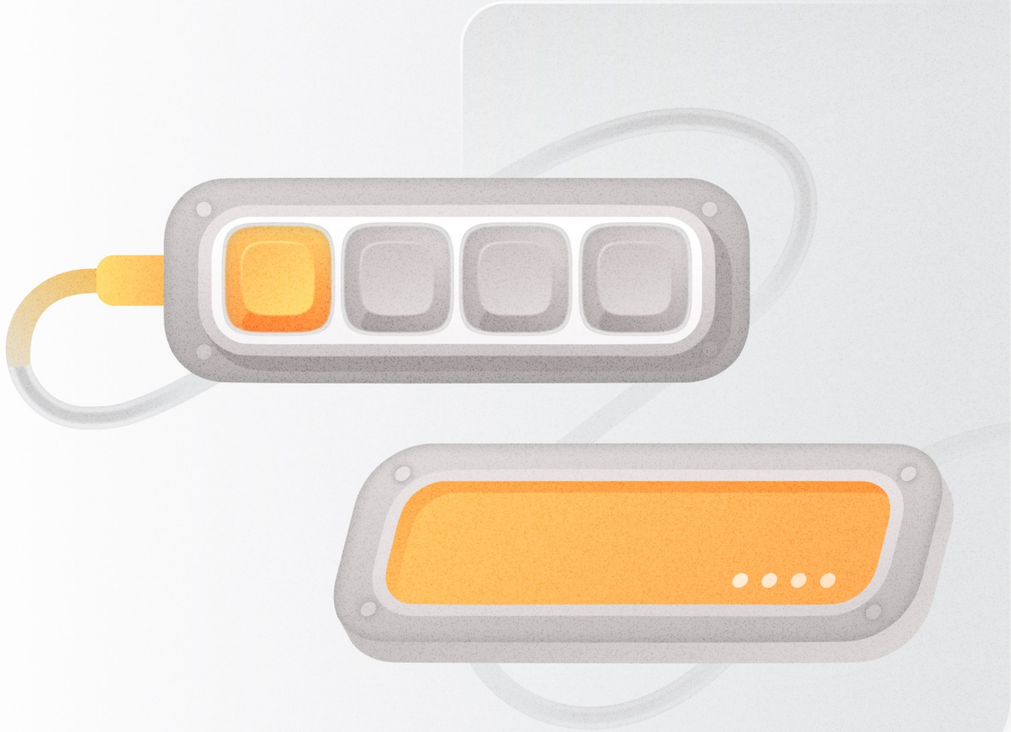
- Минималистичный образ ОС
- Ноды проектируются как временная сущность

Базовый Talos

- Вся ОС – сжатый initramfs – 88 MiB
- Rootfs – примонтированный squashfs
- Всего около 40 бинарников



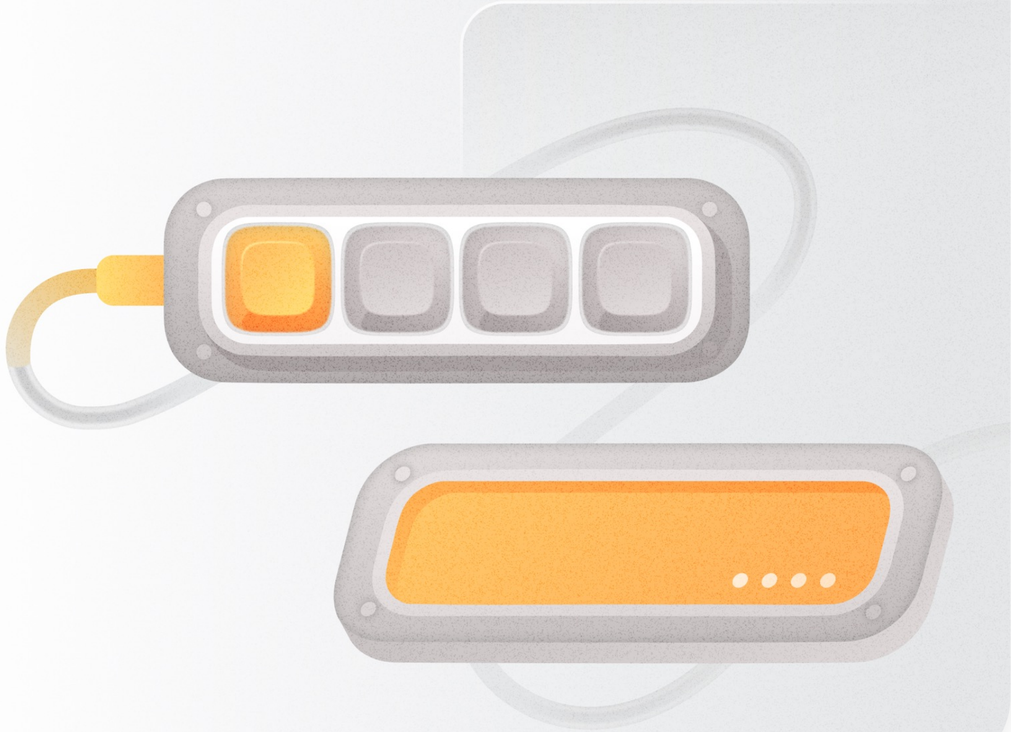
Talos и GPU: Сложности



Модули ядра

- Запрещена динамическая загрузка модулей ядра
- Для загрузки модуля ядра нужна подпись

Talos и GPU: Сложности



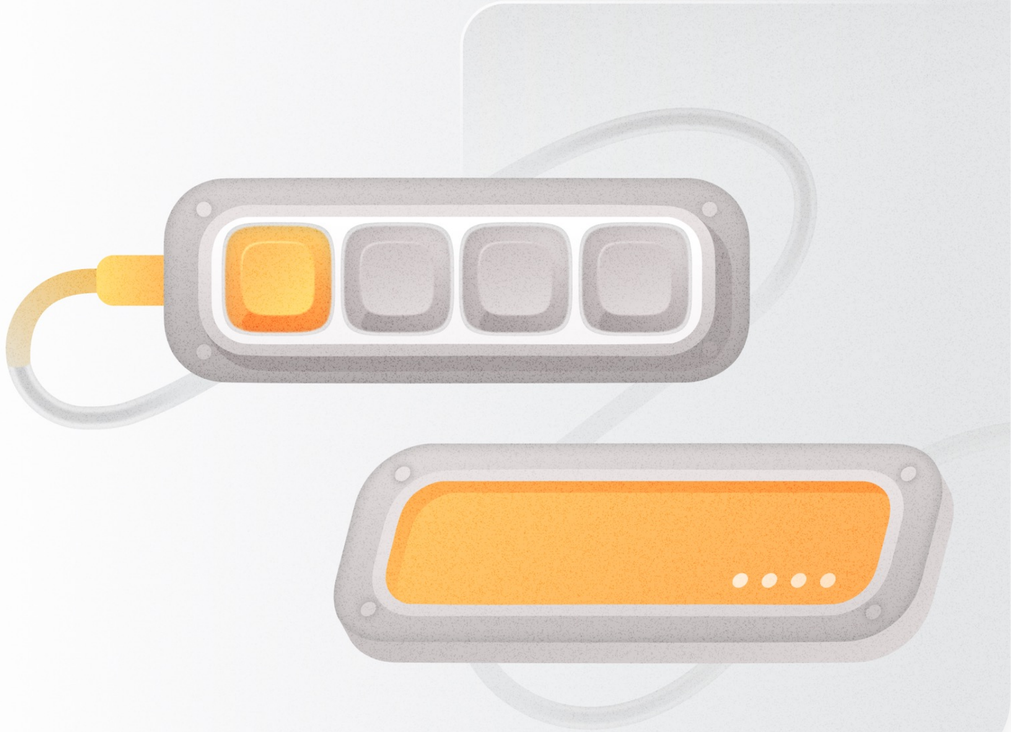
Модули ядра

- Запрещена динамическая загрузка модулей ядра
- Для загрузки модуля ядра нужна подпись

C Library

- Talos Linux использует musl
- Все компоненты NVIDIA ориентированы на glibc

Talos и GPU: Сложности



Модули ядра

- Запрещена динамическая загрузка модулей ядра
- Для загрузки модуля ядра нужна подпись

C Library

- Talos Linux использует musl
- Все компоненты NVIDIA ориентированы на glibc

Отсутствие shell

Container Toolkit использует shell обертки

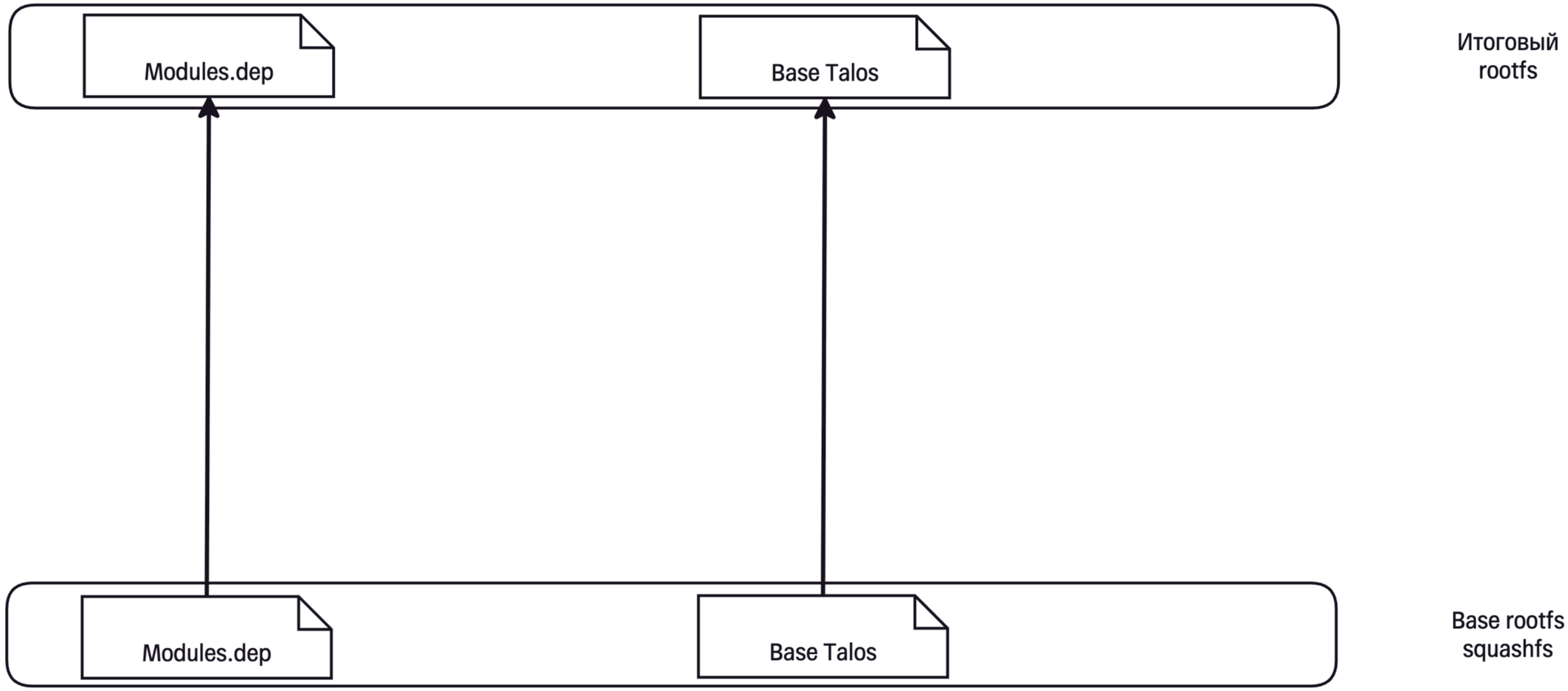
Talos и GPU: Talos <1.13



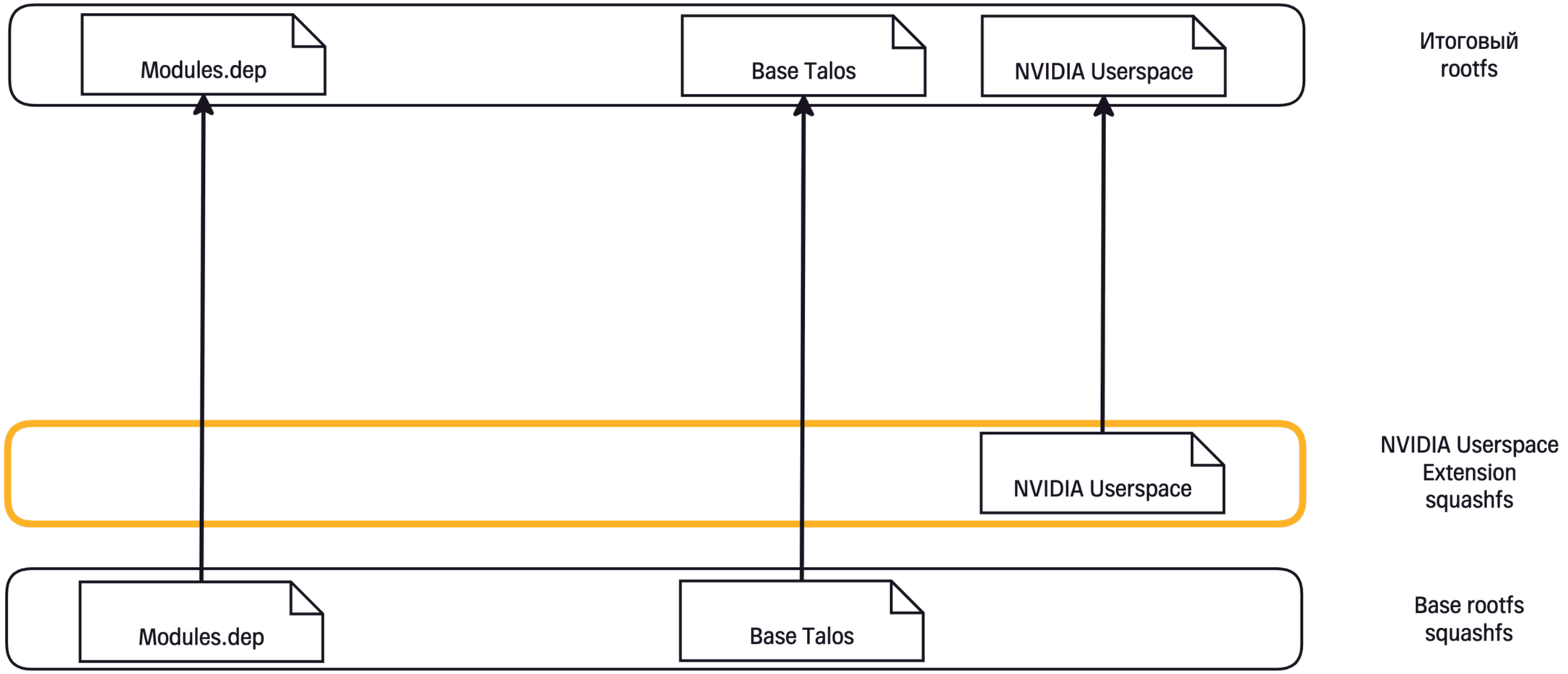
Extensions

- Модули ядра NVIDIA-драйвера
- Userspace компоненты NVIDIA – Container Toolkit и зависимости

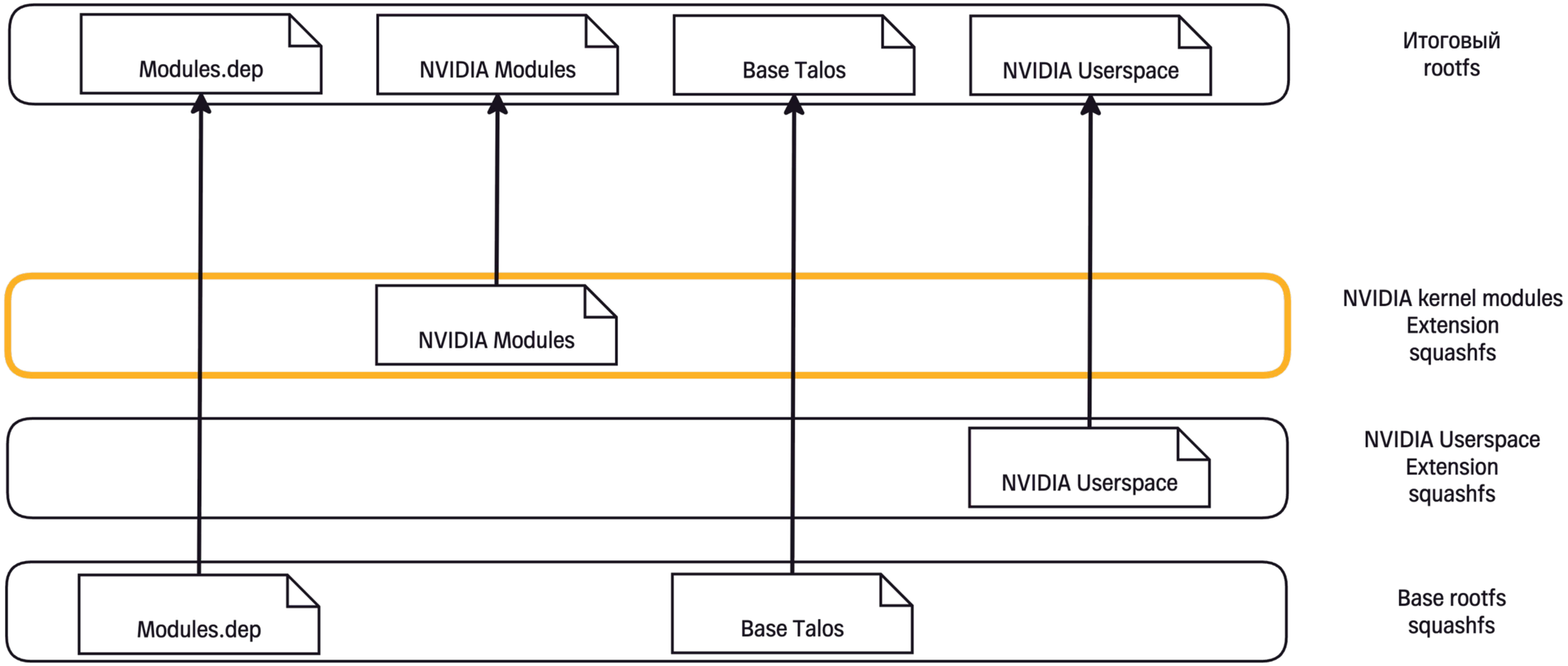
Extensions


Modules.dep
Base Talos
Итоговый
rootfs
Modules.dep
Base Talos
Base rootfs
squashfs

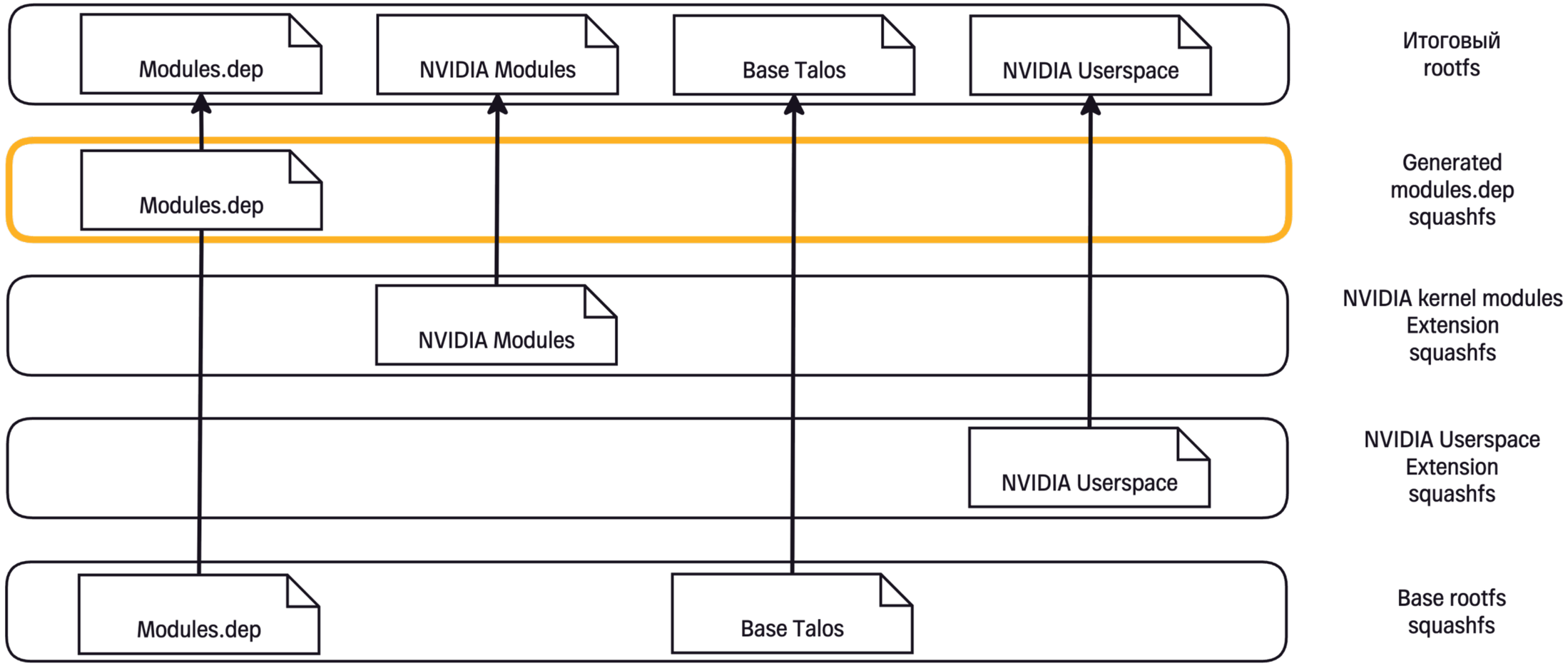
Extensions


Modules.dep
Base Talos
NVIDIA Userspace
Итоговый rootfs
NVIDIA Userspace
NVIDIA Userspace Extension squashfs
Modules.dep
Base Talos
Base rootfs squashfs

Extensions

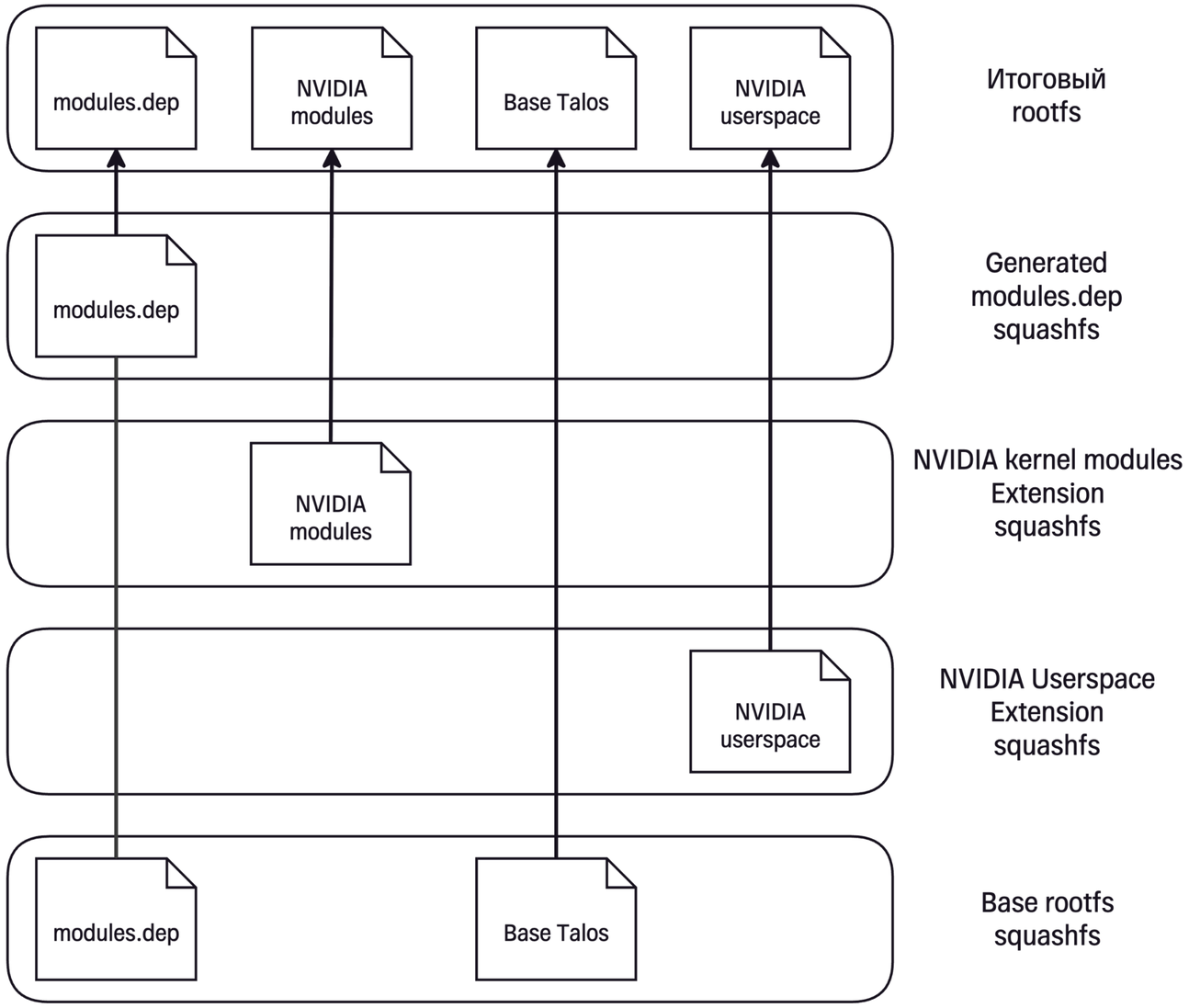

Modules.dep
NVIDIA Modules
Base Talos
NVIDIA Userspace
Итоговый rootfs
NVIDIA Modules
NVIDIA kernel modules Extension squashfs
NVIDIA Userspace
NVIDIA Userspace Extension squashfs
Modules.dep
Base Talos
Base rootfs squashfs

Extensions


Modules.dep
NVIDIA Modules
Base Talos
NVIDIA Userspace
Итоговый rootfs
Modules.dep
Generated modules.dep squashfs
NVIDIA Modules
NVIDIA kernel modules Extension squashfs
NVIDIA Userspace
NVIDIA Userspace Extension squashfs
Modules.dep
Base Talos
Base rootfs squashfs

Extensions

- Генерируется новый modules.dep
- Все squashfs зашиваются в initramfs образа ОС
- При запуске Talos extensions монтируются друг над другом по принципу overlayfs



Extensions: Особенности

Сложности с обновлением

- Версии компонентов привязаны к версии Talos
- Обновление компонентов в рамках текущей версии Talos – кастомная сборка extensions

Extensions: Особенности

Сложности с обновлением

- Версии компонентов привязаны к версии Talos
- Обновление компонентов в рамках текущей версии Talos – кастомная сборка extensions

Подпись модулей ядра

- Подпись уникальна и доступна только на этапе сборки
- Обновление версии модуля – пересборка ядра Talos

Talos и GPU: Talos <1.13



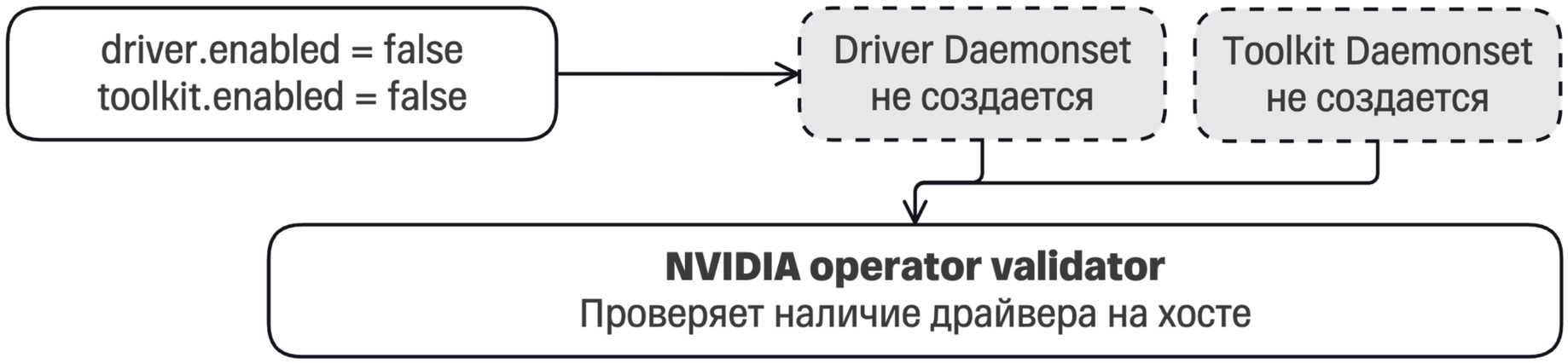
Extensions

- Модули ядра NVIDIA-драйвера
- Userspace компоненты NVIDIA – Container Toolkit и зависимости

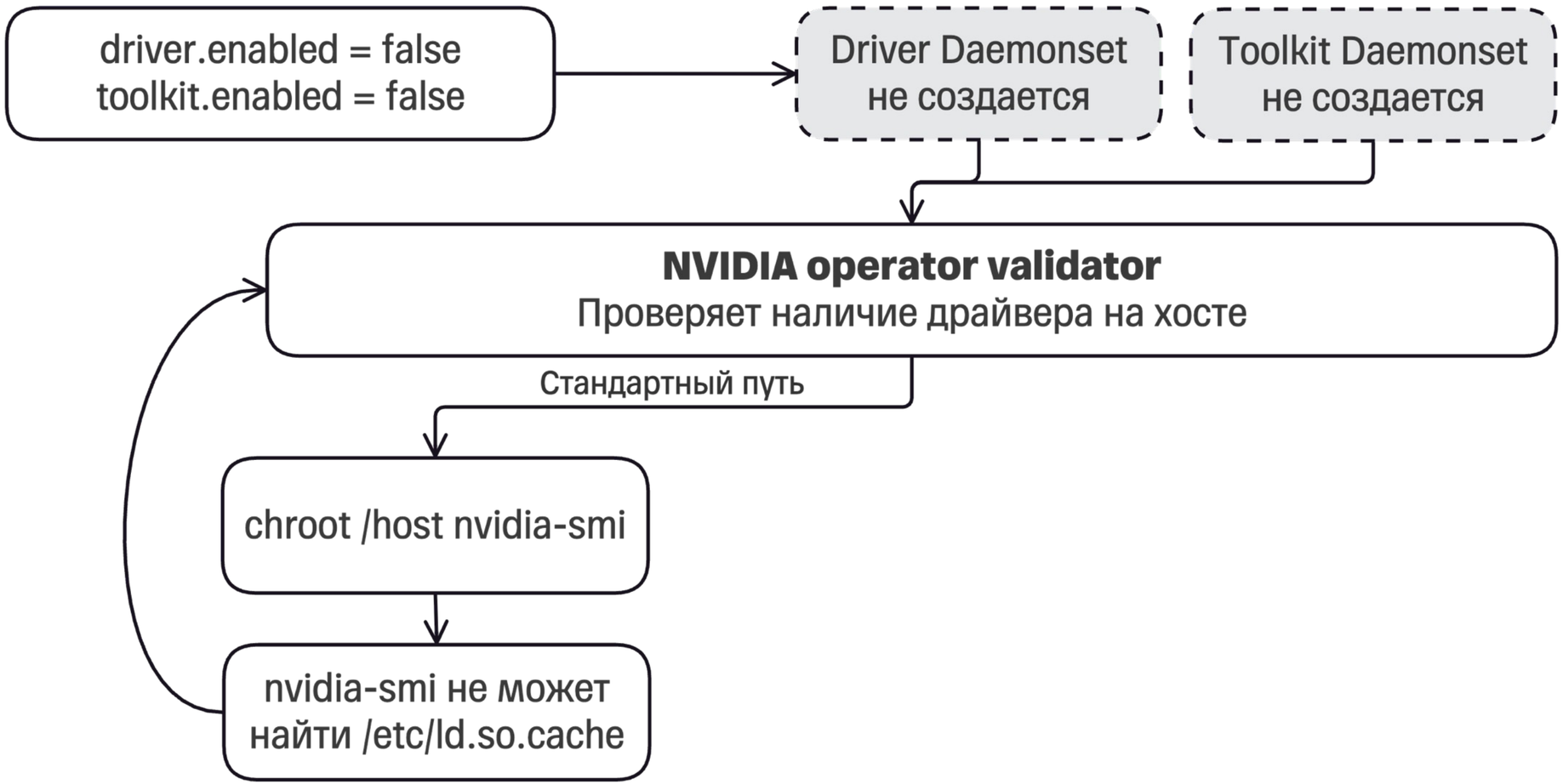


NVIDIA GPU Operator?

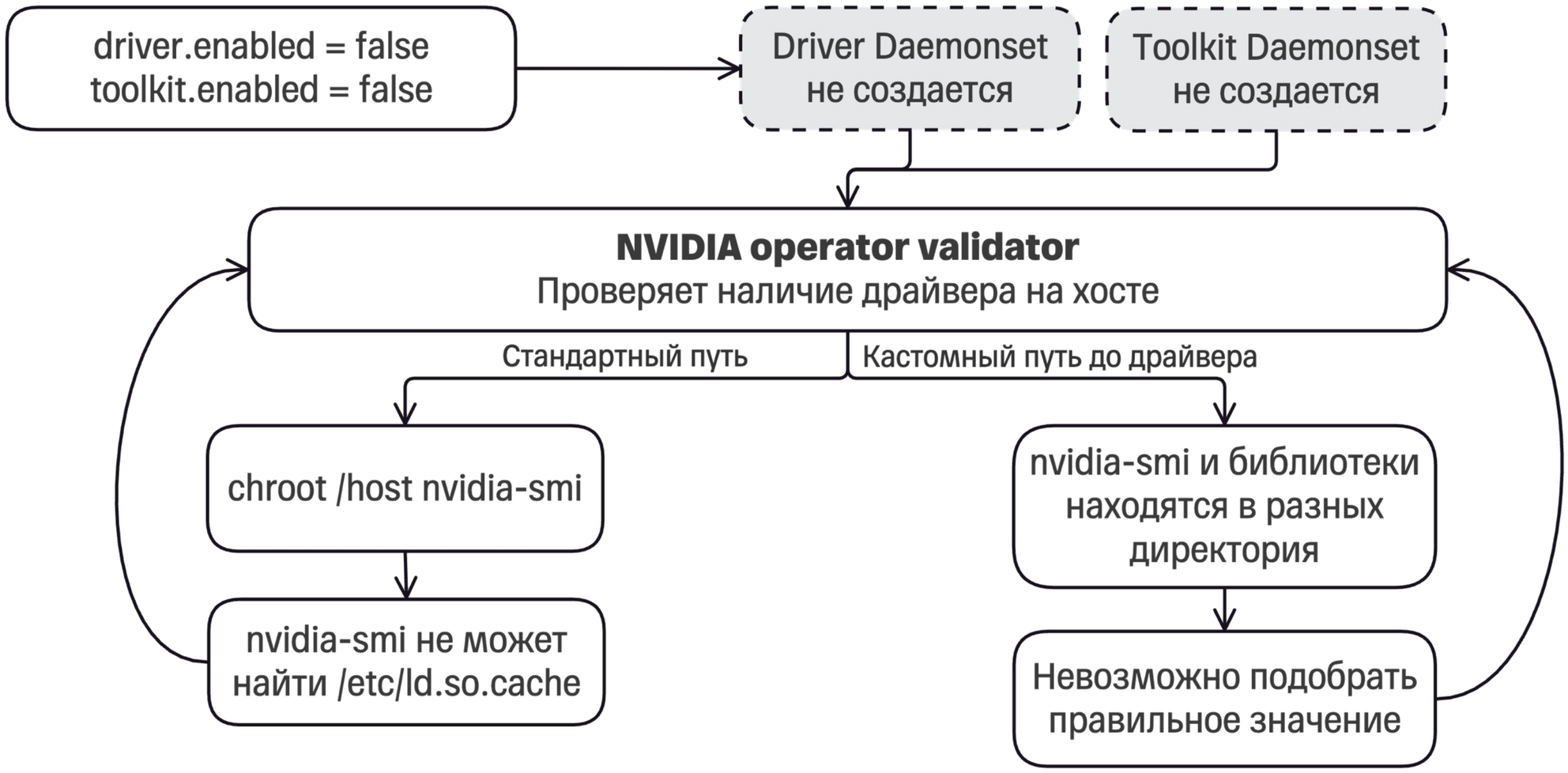
NVIDIA GPU Operator?



NVIDIA GPU Operator?



NVIDIA GPU Operator?



Talos и GPU: Talos <1.13



Extensions

- Модули ядра NVIDIA-драйвера
- Userspace компоненты NVIDIA – Container Toolkit и зависимости



Компоненты GPU Operator

- NVIDIA Device Plugin
- Мониторинг

Talos и GPU: Промежуточные итоги



Вышли в Production

Запустили два production кластера

Talos и GPU: Промежуточные итоги



Вышли в Production

Запустили два production кластера



Уязвимости

- Получили две CVE
- Проблема решилась обновлением Container toolkit

Talos и GPU: Промежуточные итоги

Вышли в Production

Запустили два production кластера

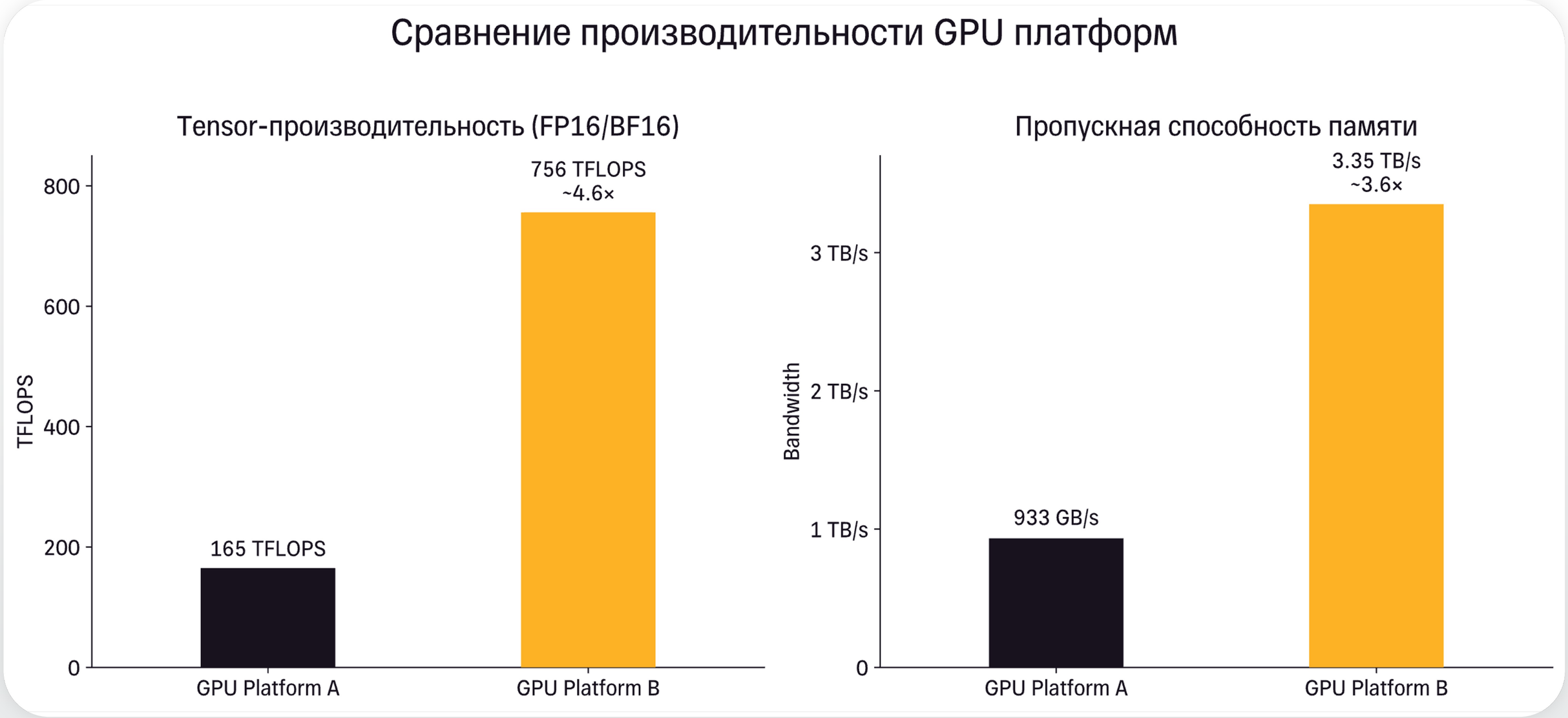
Уязвимости

- Получили две CVE
- Проблема решилась обновлением Container toolkit

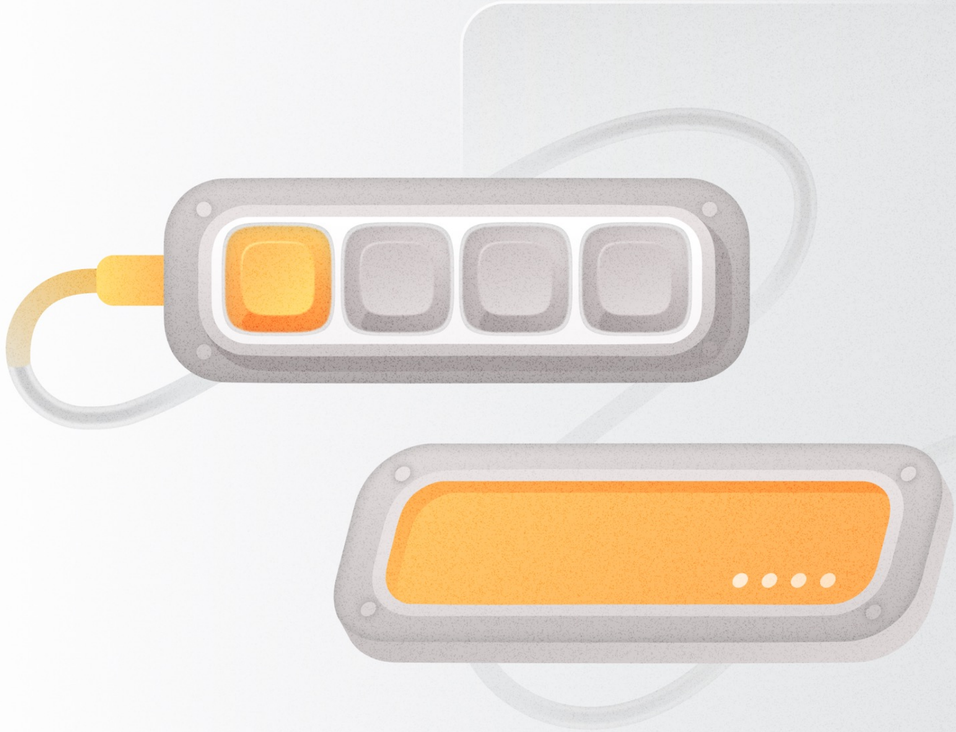
Развитие LLM

Используемая GPU платформа не позволяла запускать LLM

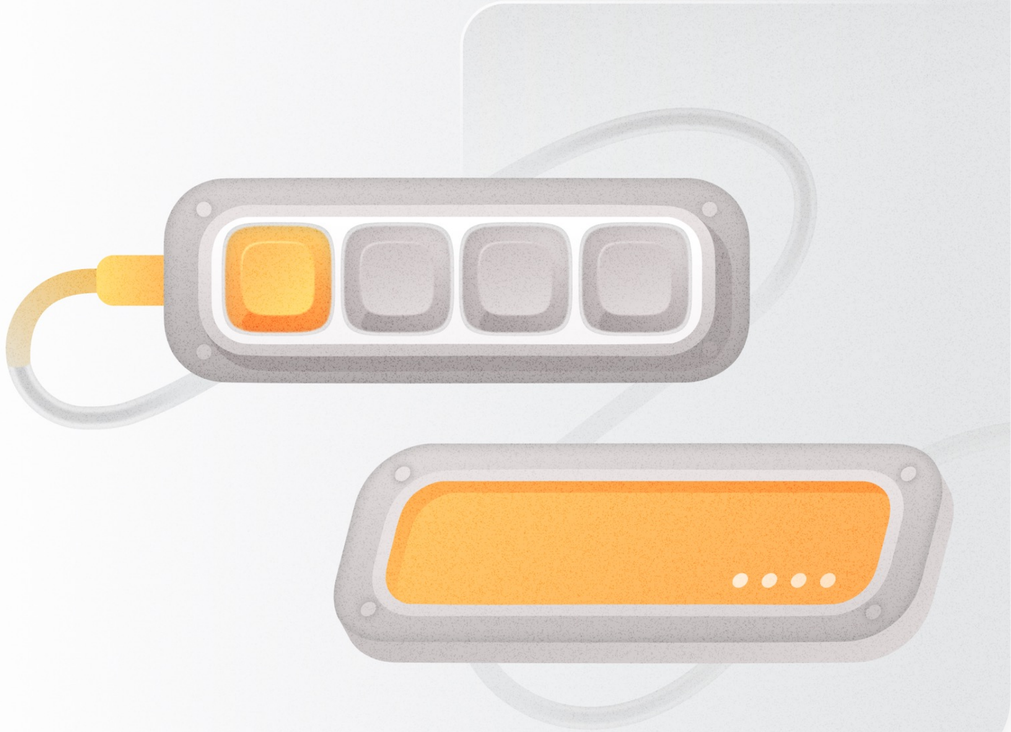
Переход к новой GPU платформе



Проблема 0
Запуск первого кластера



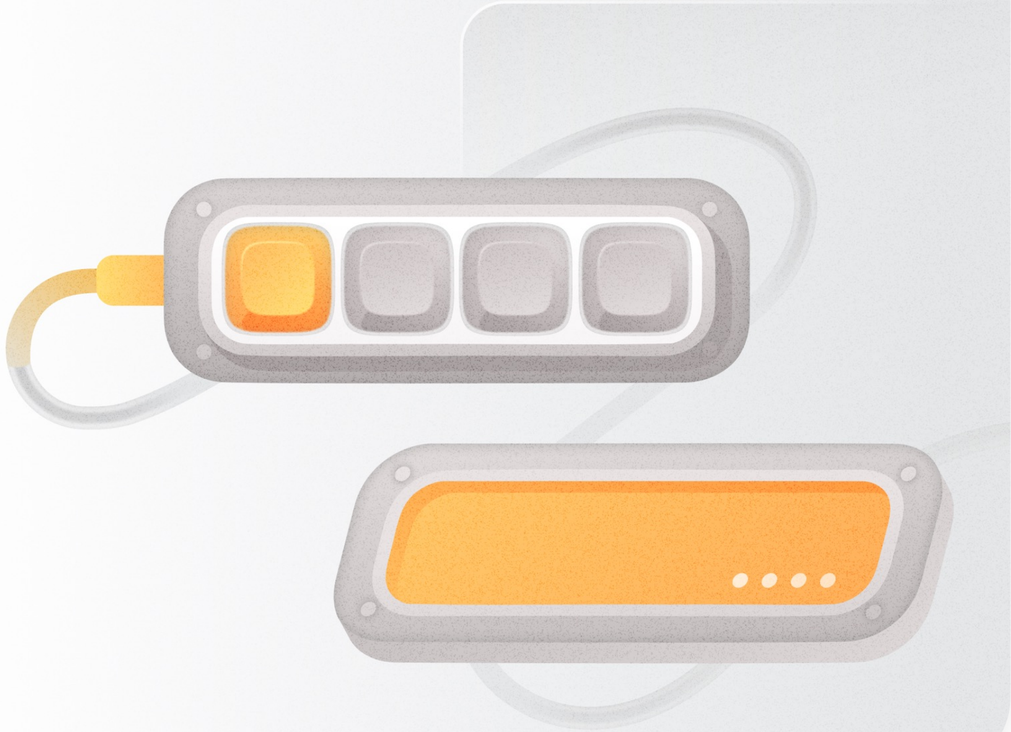
Проблема 0
Запуск первого кластера



Подготовили Talos

На ноды устанавливается Talos с NVIDIA GPU Extensions

Проблема 0
Запуск первого кластера



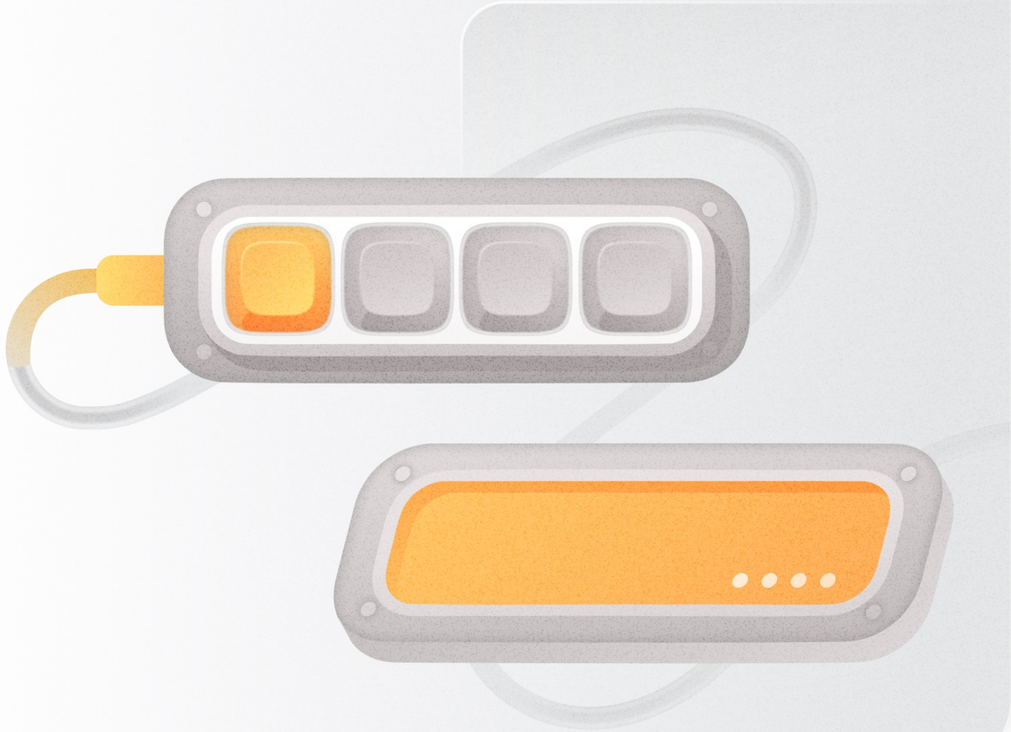
Подготовили Talos

На ноды устанавливается Talos с NVIDIA GPU Extensions

K8S успешно определил GPU

Видеокарты видны в ресурсах ноды

Проблема 0
Запуск первого кластера



Подготовили Talos

На ноды устанавливается Talos с NVIDIA GPU Extensions

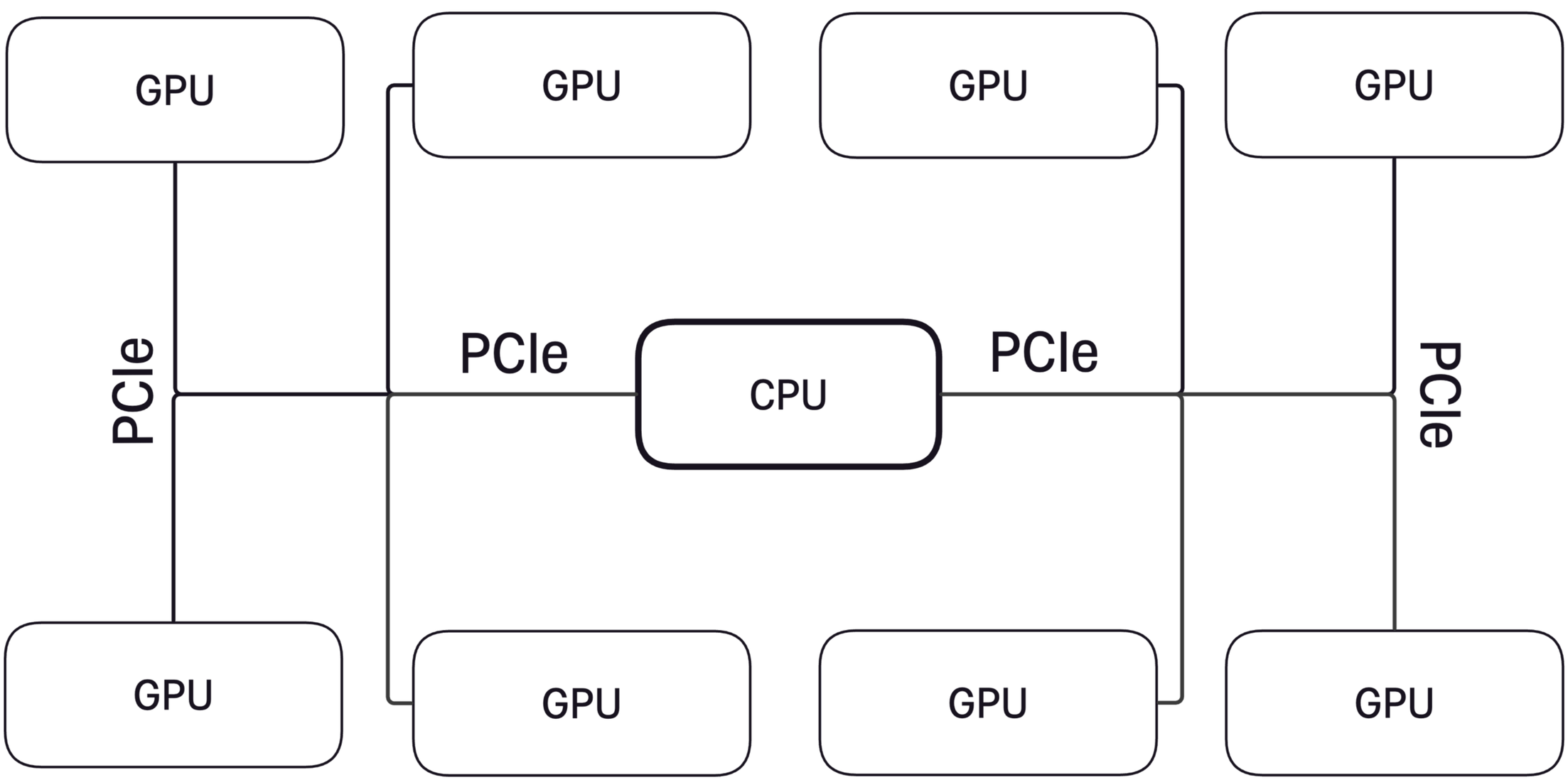
K8S успешно определил GPU

Видеокарты видны в ресурсах ноды

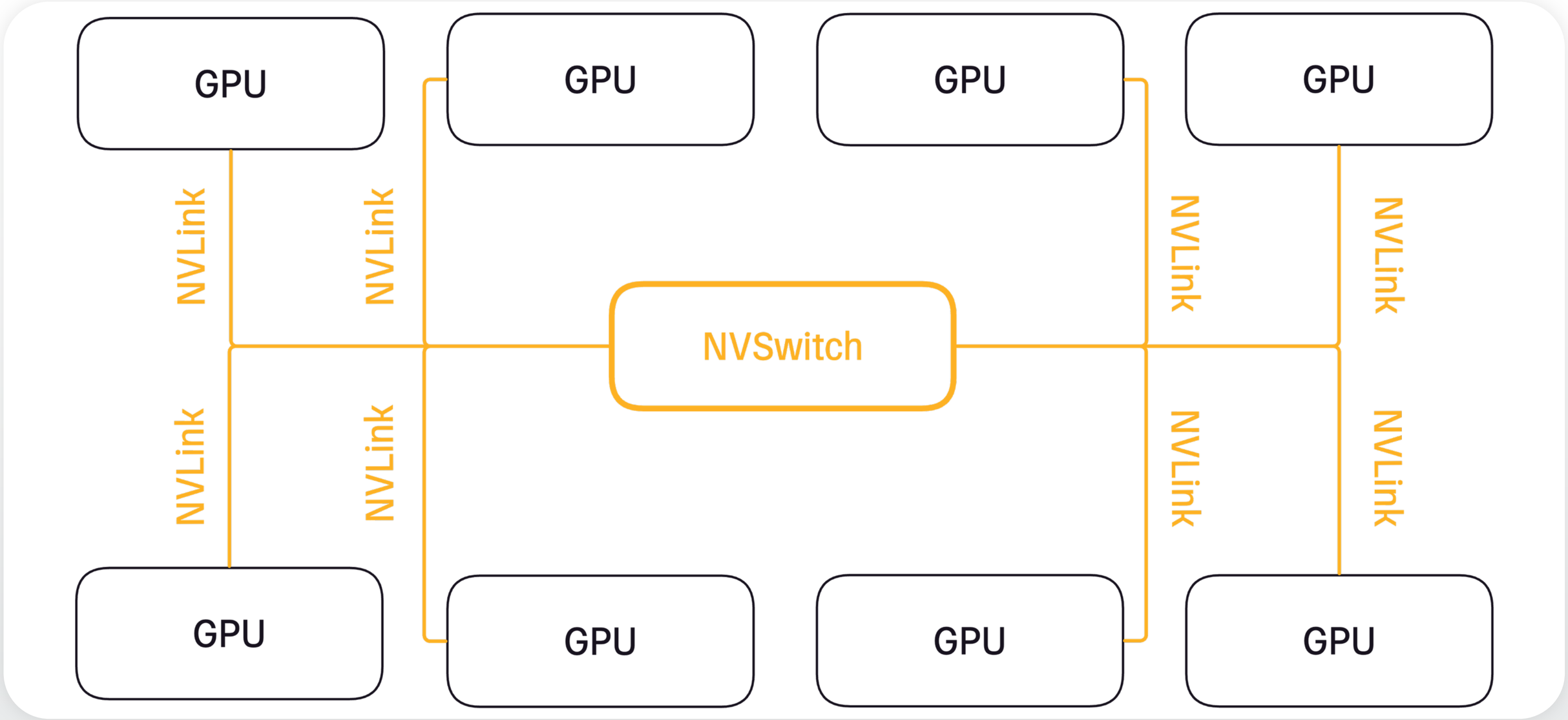
Нагрузка не запускается

GPU нагрузка отдает ошибку, что система непроинициализирована

Сервер с GPU: Ожидание

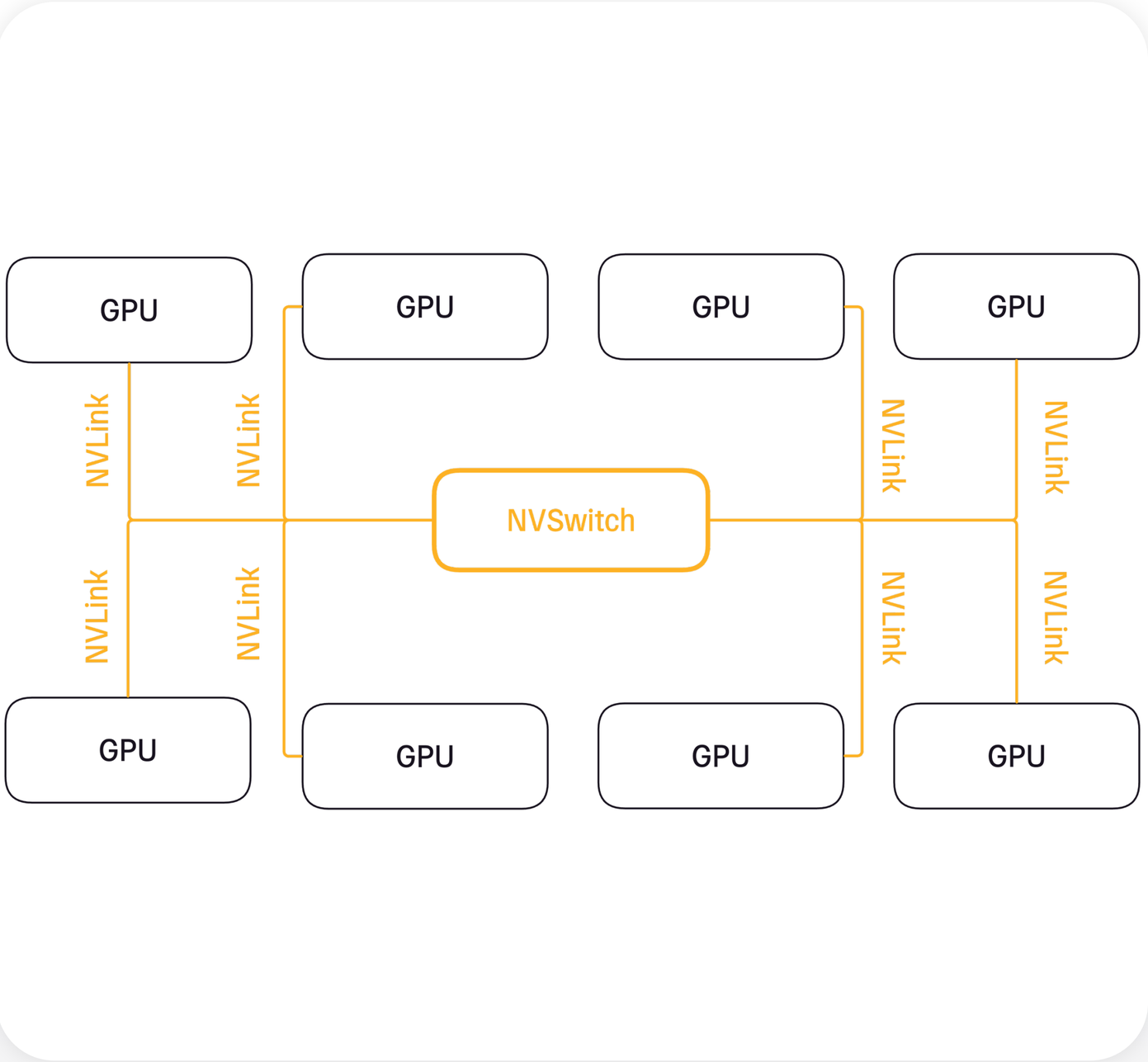


Сервер с GPU: Реальность



NVSwitch фабрика

- GPU подключены в NVSwitch фабрику через NVLink
- NVLink быстрее PCIe в 7 раз
- Без инициализации NVSwitch драйвер не позволяет запускать нагрузки



NVIDIA Fabric Manager

Отдельный демон

- Отвечает за инициализацию NVSwitch фабрики
- Настраивает маршрутизацию и топологию кластера

NVIDIA Fabric Manager

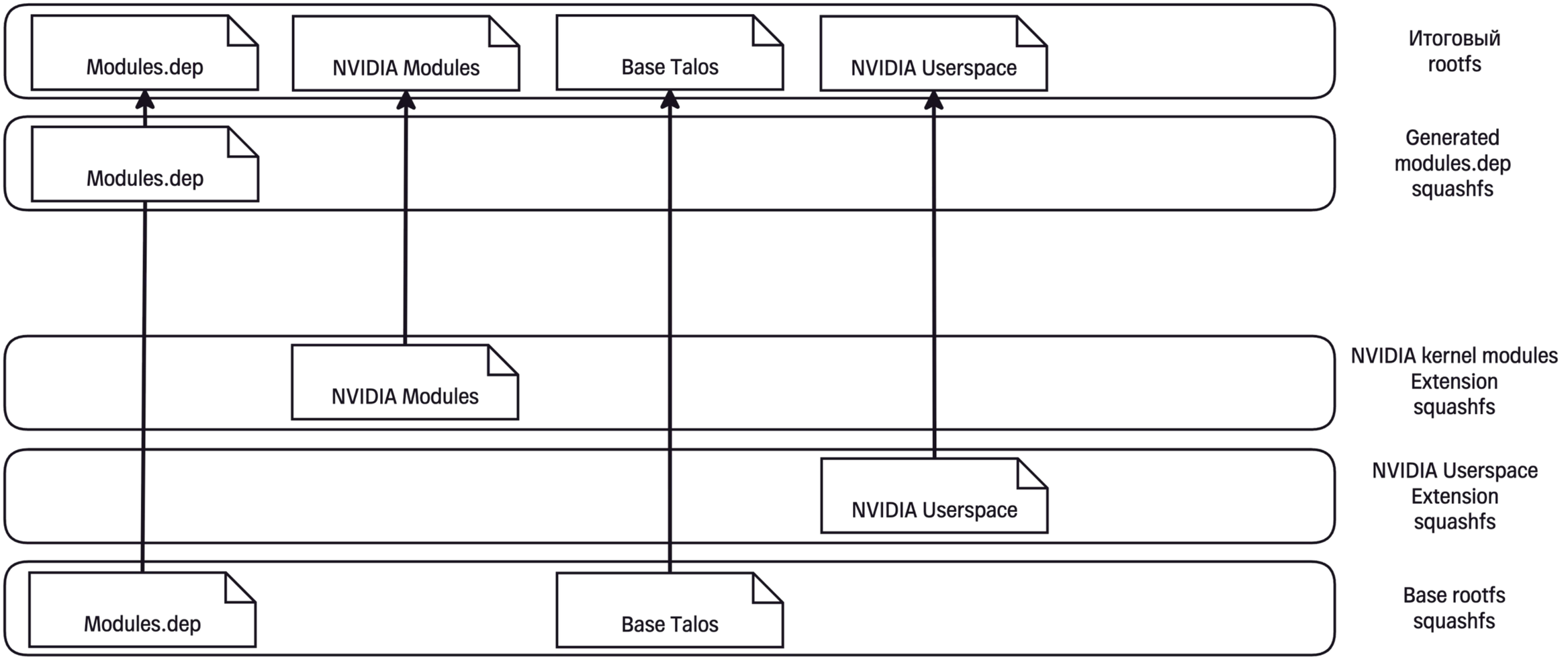
Отдельный демон

- Отвечает за инициализацию NVSwitch фабрики
- Настраивает маршрутизацию и топологию кластера

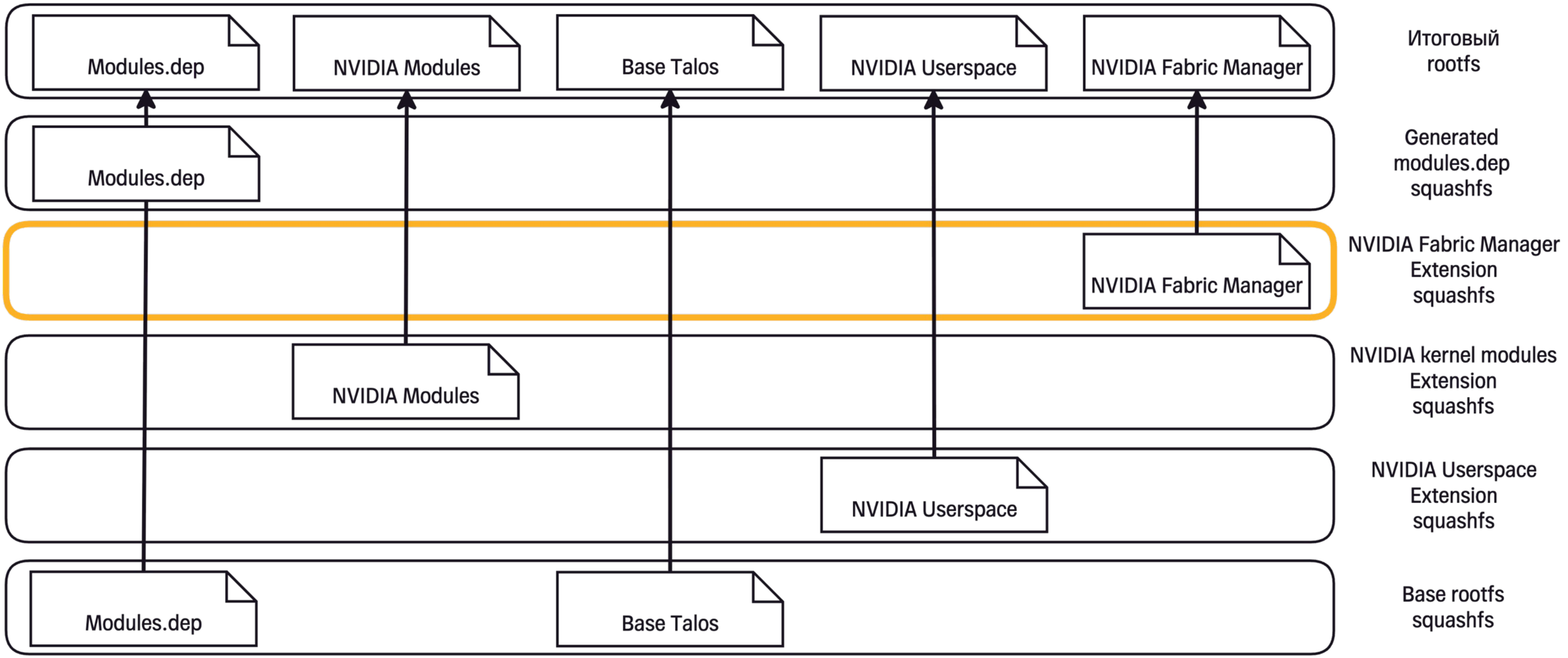
Talos extension

- Отдельный extension для установки NVIDIA Fabric Manager
- В opensource отсутствовала топология для нашего GPU

Talos Extension



Talos Extension



Промежуточные выводы



Новая сборка Talos

Подготовили новую сборку Talos специально под новую GPU платформу

Промежуточные выводы



Новая сборка Talos

Подготовили новую сборку Talos специально под новую GPU платформу



Тестирование

Расширили тестирование. Добавили запуск пробной GPU нагрузки

Промежуточные выводы



Новая сборка Talos

Подготовили новую сборку Talos специально под новую GPU платформу



Тестирование

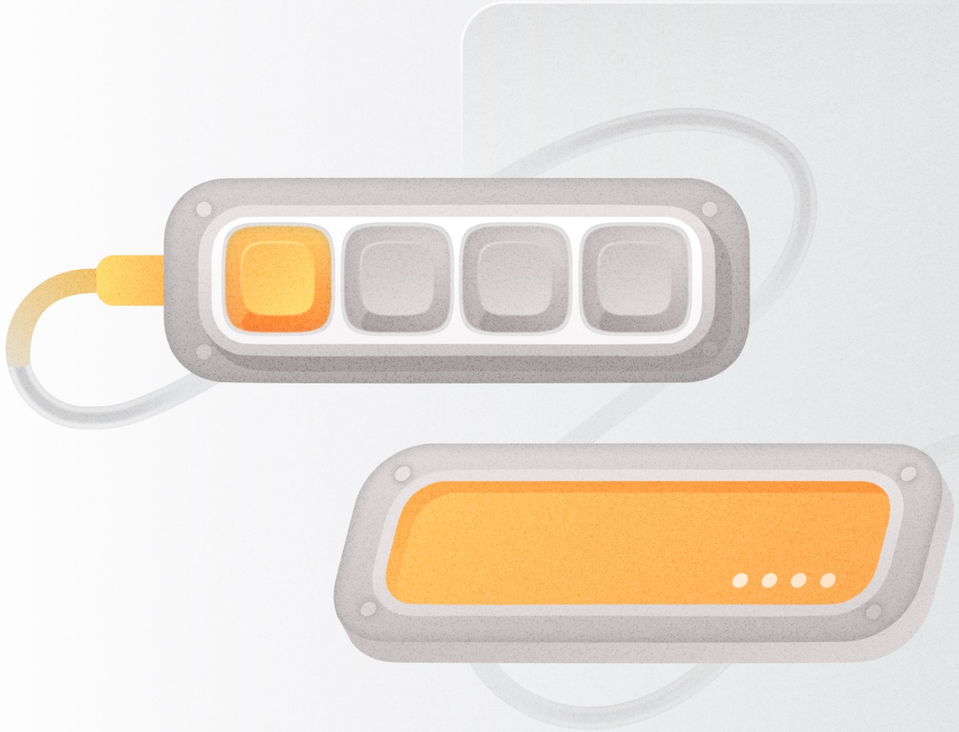
Расширили тестирование. Добавили запуск пробной GPU нагрузки



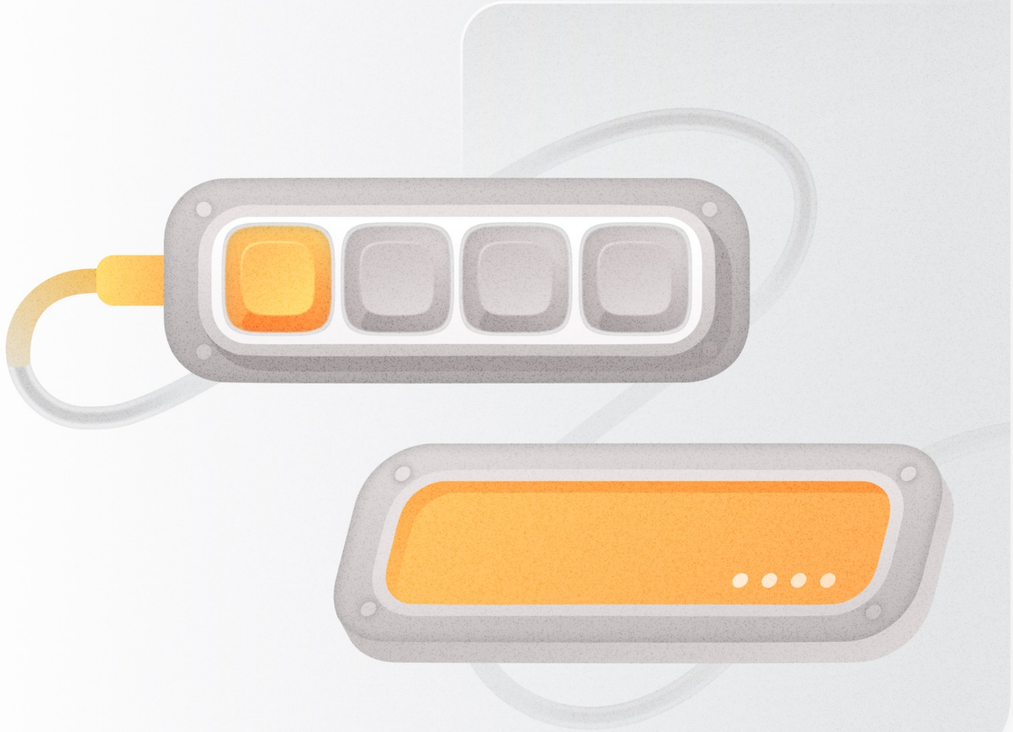
Запустили кластер

Пользователи получили возможность использовать LLM внутри банка

Запуск второго кластера



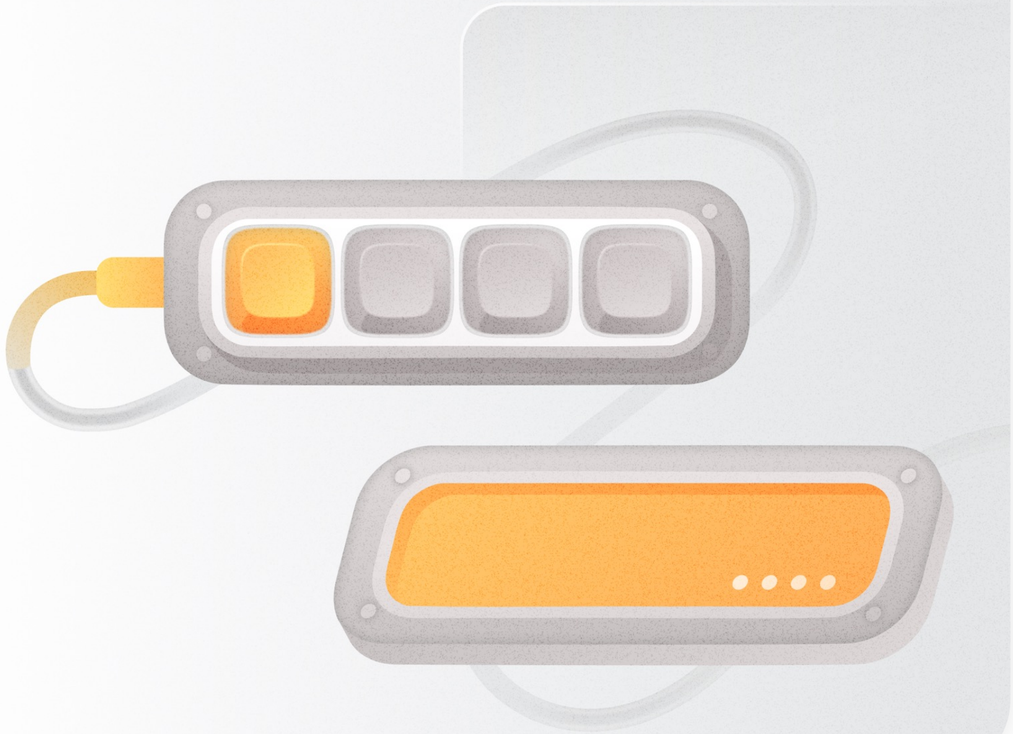
Запуск второго кластера



Новый вендор

Опыта работы с этим вендором до этого не было

Запуск второго кластера



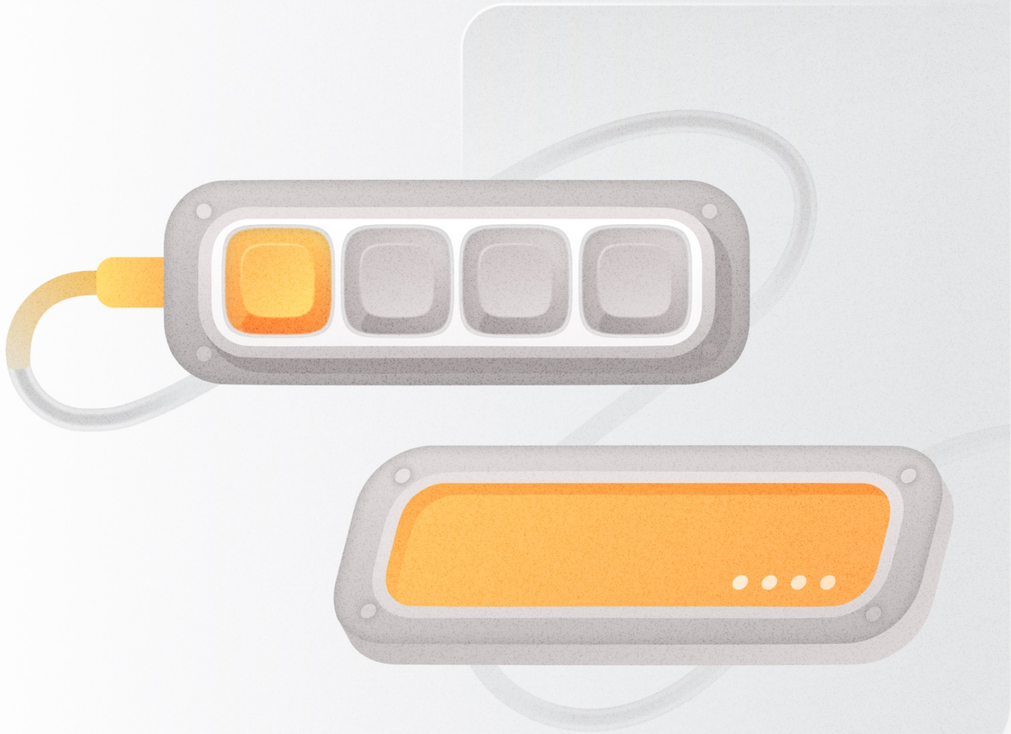
Новый вендор

Опыта работы с этим вендором до этого не было

Новый Talos

Большое количество компонентов обновляется

Запуск второго кластера



Новый вендор

Опыта работы с этим вендором до этого не было

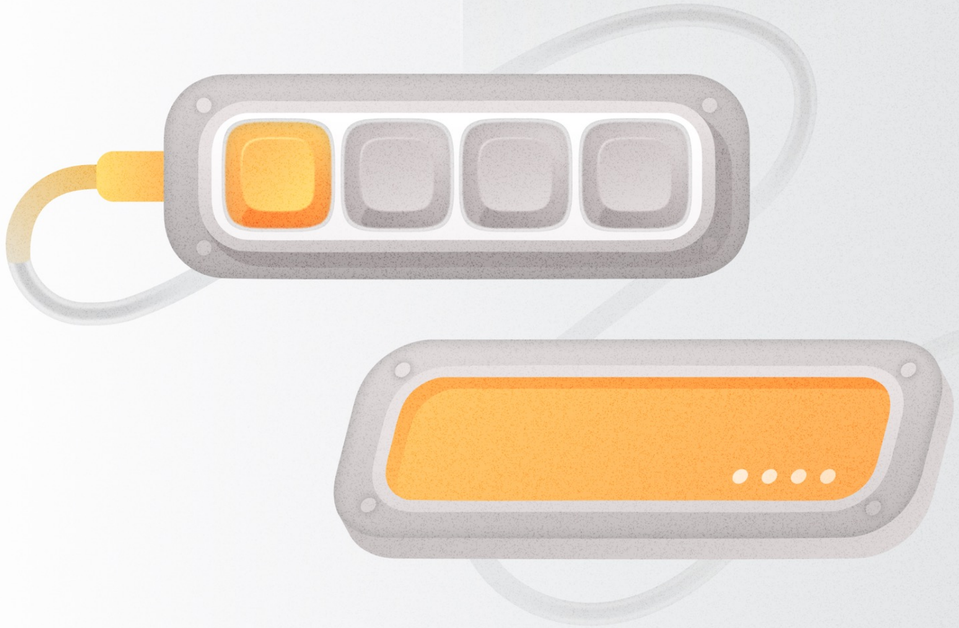
Новый Talos

Большое количество компонентов обновляется

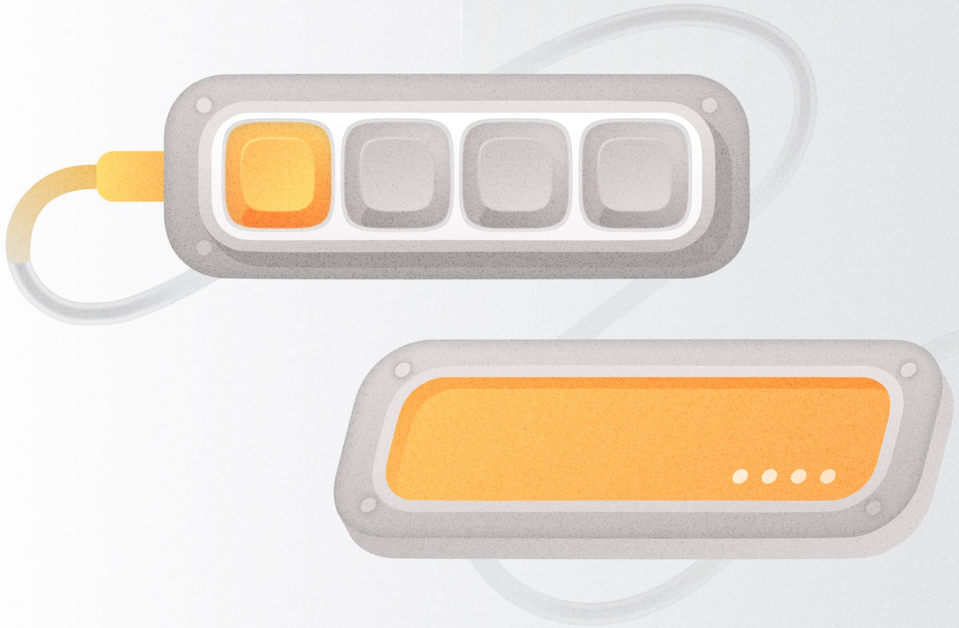
Cgroups v2

Пользовательские приложения долго не могли перейти с cgroupsv1

Проблема 1
Медленный запуск контейнеров



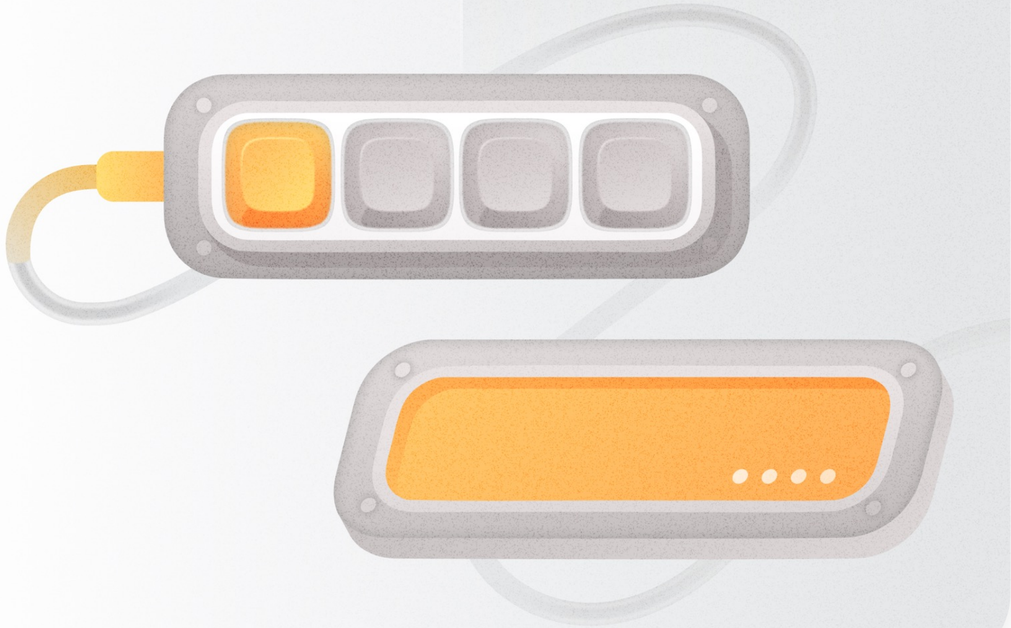
Проблема 1
Медленный запуск контейнеров



Подготовили Talos

- На ноды устанавливается Talos с NVIDIA GPU Extensions
- Дополнительно запущены проверки

Проблема 1 Медленный запуск контейнеров



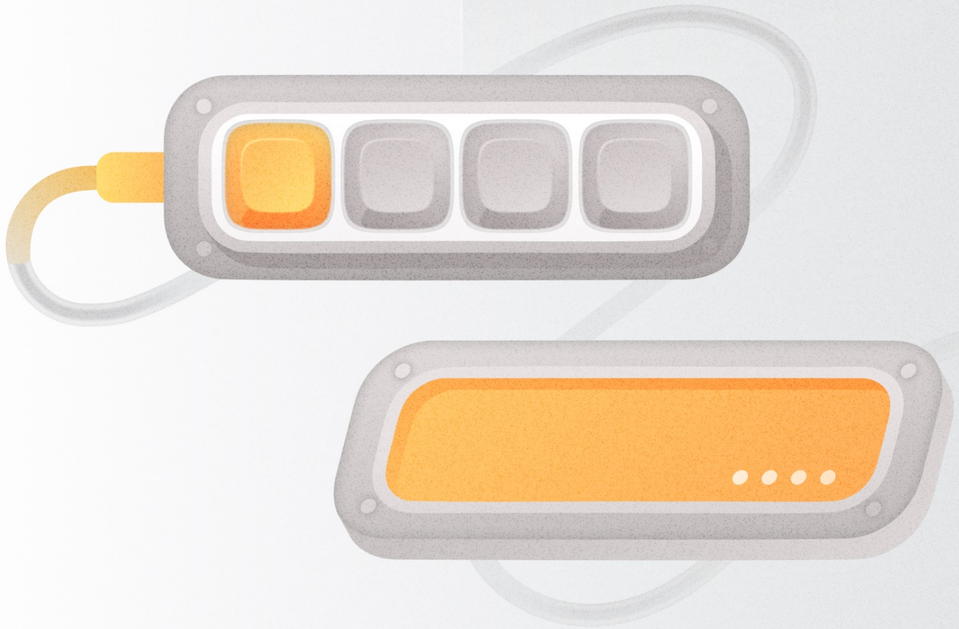
Подготовили Talos

- На ноды устанавливается Talos с NVIDIA GPU Extensions
- Дополнительно запущены проверки

Пользователи жалуются

Контейнеры долго висят в ContainerCreating

Проблема 1
Медленный запуск контейнеров



Подготовили Talos

- На ноды устанавливается Talos с NVIDIA GPU Extensions
- Дополнительно запущены проверки

Пользователи жалуются

Контейнеры долго висят в ContainerCreating

Начинаем расследование

- Наши тесты подтверждают проблему
- На stage проблема не воспроизводится

Ожидание

kubectl describe ×

```
Events:
  Type    Reason   Age   From     Message
  ----    ------   ----  ----     -------
  Normal  Pulled   2s    kubelet  spec.containers{cuda-vector-add}: Container image ”docker-registry.com/cuda-vector-add:cuda12.5.0" already present on machine
  Normal  Created  2s    kubelet  spec.containers{cuda-vector-add}: Created container: cuda-vector-add
  Normal  Started  1s    kubelet  spec.containers{cuda-vector-add}: Started container cuda-vector-add
```

Ожидание

kubectl describe ×

```
Events:
  Type    Reason   Age   From     Message
  ----    ------   ----  ----     -------
  Normal  Pulled   2s    kubelet  spec.containers{cuda-vector-add}: Container image ”docker-registry.com/cuda-vector-add:cuda12.5.0" already present on machine
  Normal  Created  2s    kubelet  spec.containers{cuda-vector-add}: Created container: cuda-vector-add
  Normal  Started  1s    kubelet  spec.containers{cuda-vector-add}: Started container cuda-vector-add
```

Быстрый переход из Created в Started

Реальность

kubectl describe ×

```
Events:
  Type    Reason   Age   From     Message
  ----    ------   ----  ----     -------
  Normal  Pulled   30s   kubelet  spec.containers{cuda-vector-add}: Container image "docker-
registry.com/cuda-vector-add:cuda12.5.0" already present on machine
  Normal  Created  30s   kubelet  spec.containers{cuda-vector-add}: Created container: cuda-
vector-add
  Normal  Started   1s   kubelet  spec.containers{cuda-vector-add}: Started container cuda-
vector-add
```

Долгий переход из Created в Started

30 секунд - много ли это?

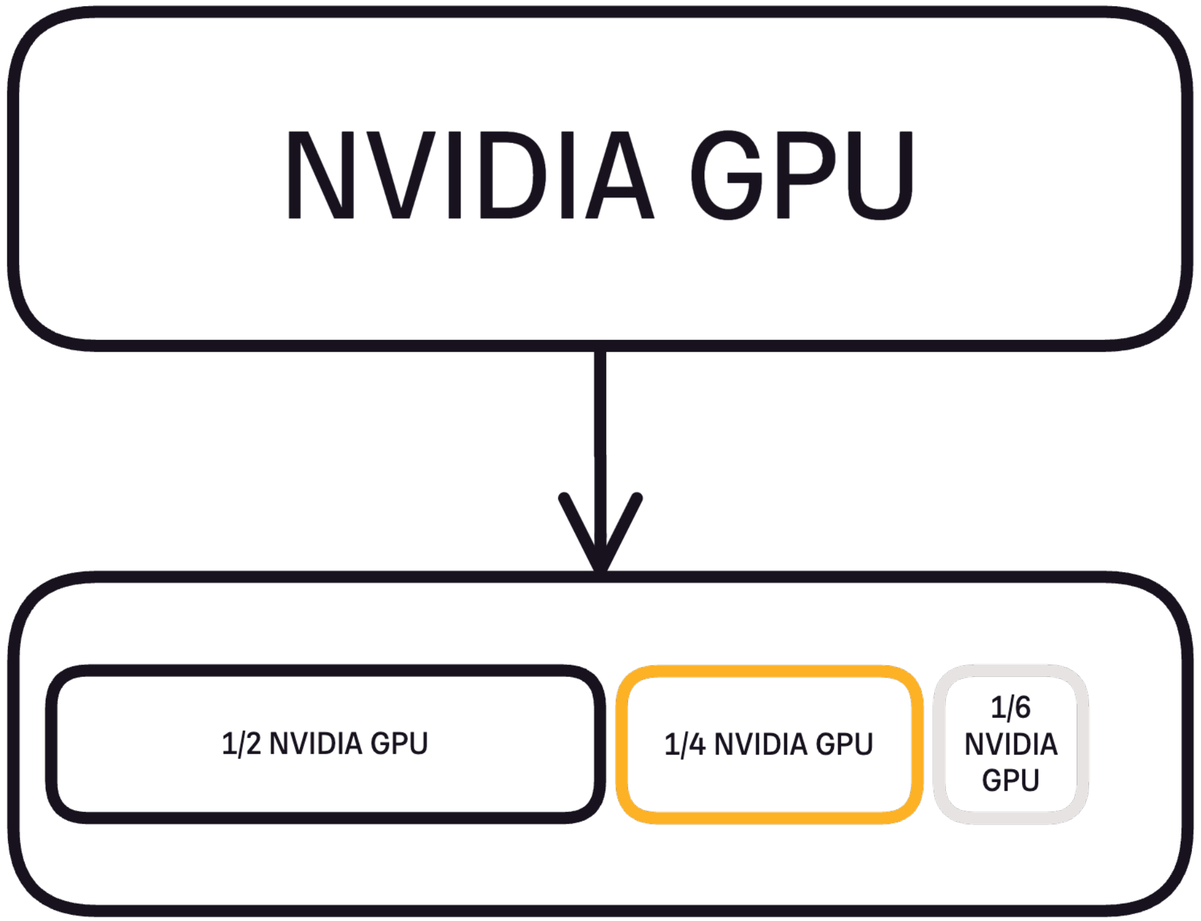


ML нагрузка запускается достаточно долго:

- Загрузка модели – от 5 минут
- Инициализация весов – от 10 минут

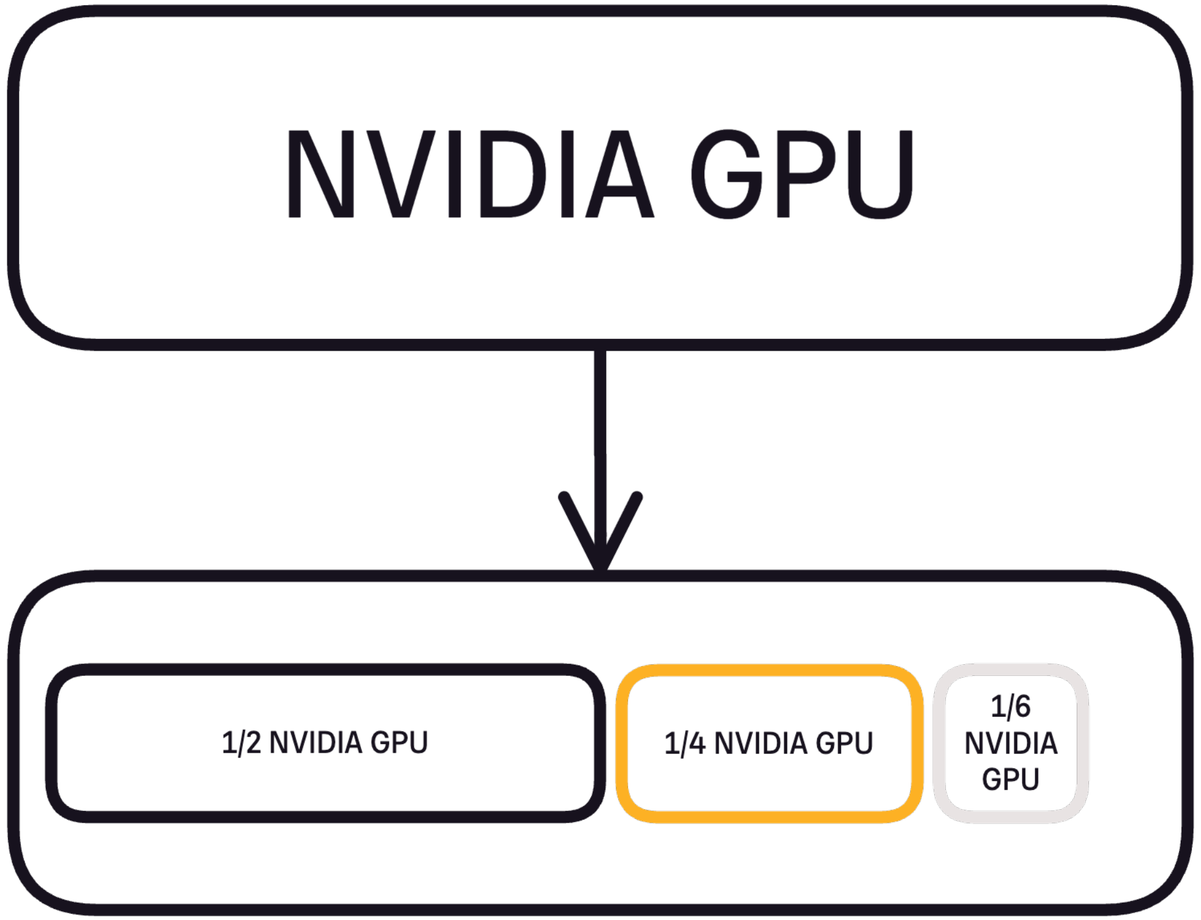
MIG– Multi-instance GPU

- Разделение GPU на несколько независимых виртуальных GPU



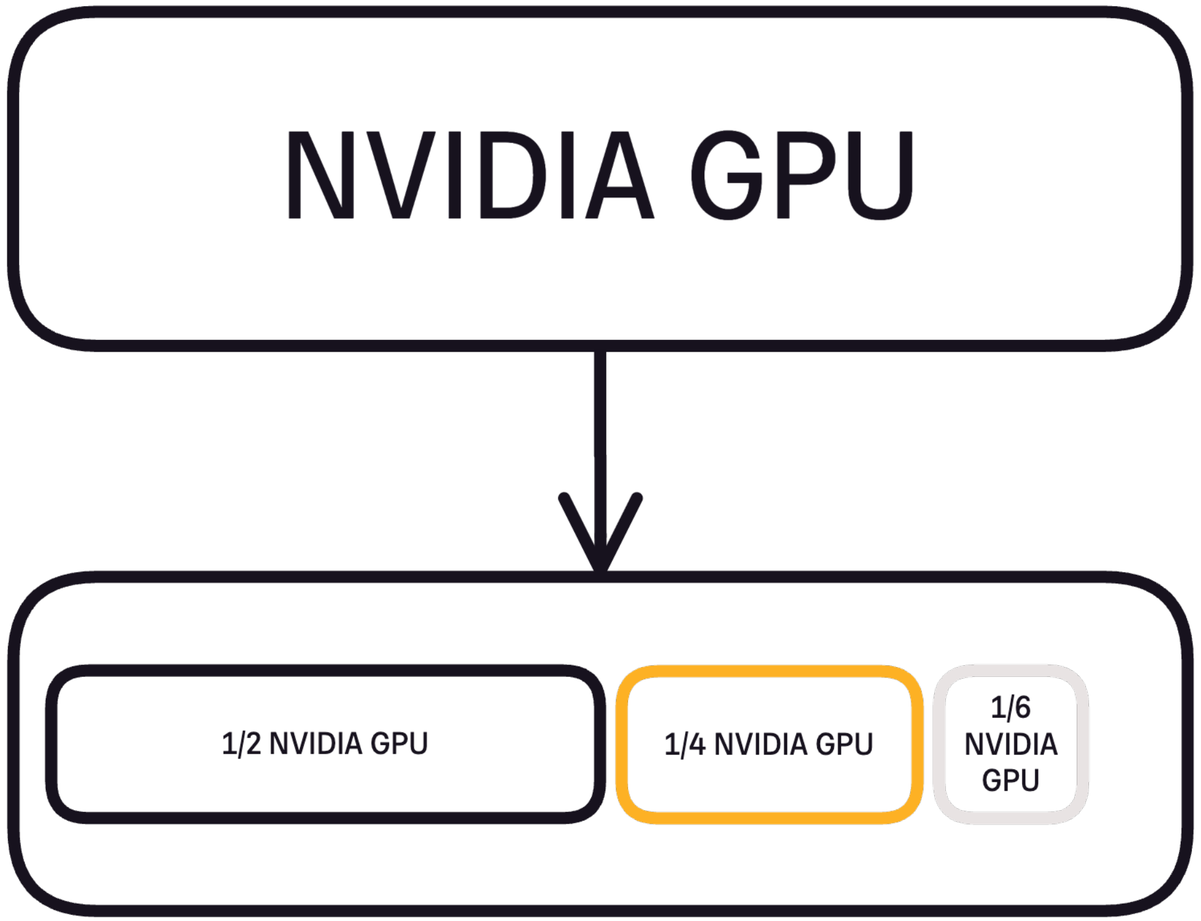
MIG– Multi-instance GPU

- Разделение GPU на несколько независимых виртуальных GPU
- Улучшается утилизация GPU



MIG– Multi-instance GPU

- Разделение GPU на несколько независимых виртуальных GPU
- Улучшается утилизация GPU
- GPU становится доступен для большего числа команд



MIG – 56 мини-GPU на ноду

kubectl describe ×

```
Events:
  Type     Reason     Age     From     Message
  ----     ------     ----    ----     -------
  Normal   Pulled     2m1s    kubelet  spec.containers{cuda-vector-add}: Container image
"docker-registry.com/cuda-vector-add:cuda12.5.0" already present on machine
  Normal   Created    2m1s    kubelet  spec.containers{cuda-vector-add}: Created container:
cuda-vector-add
  Warning  Failed     1s      kubelet  Error: failed to start container "cuda-vector-add": rpc
error: code = DeadlineExceeded desc = context deadline exceeded
```

Контейнер не успевает запуститься

Поиск проблемы

Диск

01

Поиск проблемы

Диск

Проблема проявляется именно при запуске контейнера

01

Поиск проблемы

Диск

Проблема проявляется именно при запуске контейнера

01

Cgroups V2

02

Поиск проблемы

Диск

Проблема проявляется именно при запуске контейнера

01

Cgroups V2

Длительное время пилотировали на dev кластере

02

Поиск проблемы

Диск

Проблема проявляется именно при запуске контейнера

01

Cgroups V2

Длительное время пилотировали на dev кластере

02

Containerd

03

Поиск проблемы

Диск

Проблема проявляется именно при запуске контейнера

01

Cgroups V2

Длительное время пилотировали на dev кластере

02

Containerd

Не все поды испытывают проблемы

03

Поиск проблемы

Диск

Проблема проявляется именно при запуске контейнера

01

Cgroups V2

Длительное время пилотировали на dev кластере

02

Containerd

Не все поды испытывают проблемы

03

Cgroupsv2 + NVIDIA Container Runtime

04

Поиск проблемы

Диск

Проблема проявляется именно при запуске контейнера

01

Cgroups V2

Длительное время пилотировали на dev кластере

02

Containerd

Не все поды испытывают проблемы

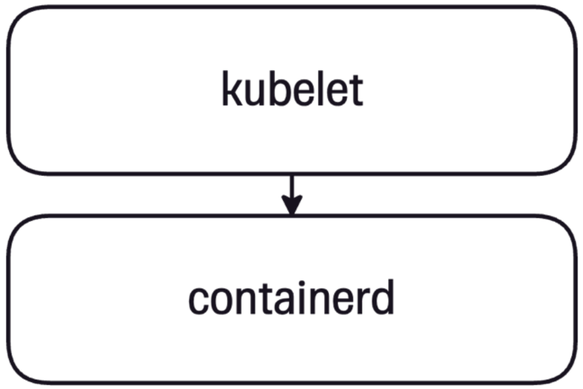
03

Cgroups v2 + NVIDIA Container runtime

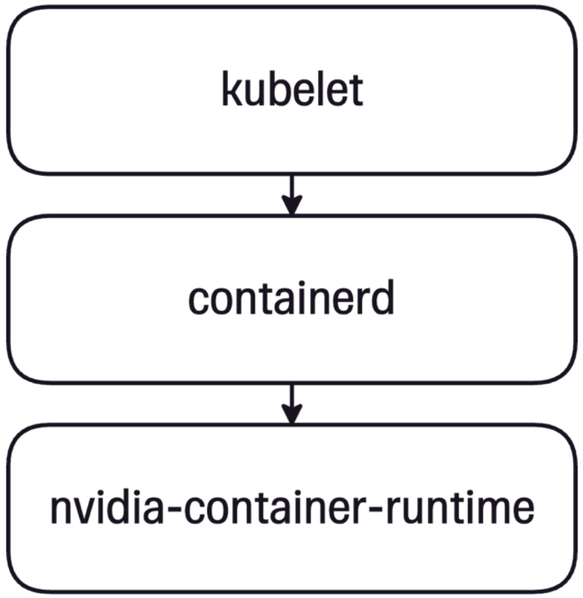
Проблему испытывают только поды, использующие NVIDIA Container Runtime

04

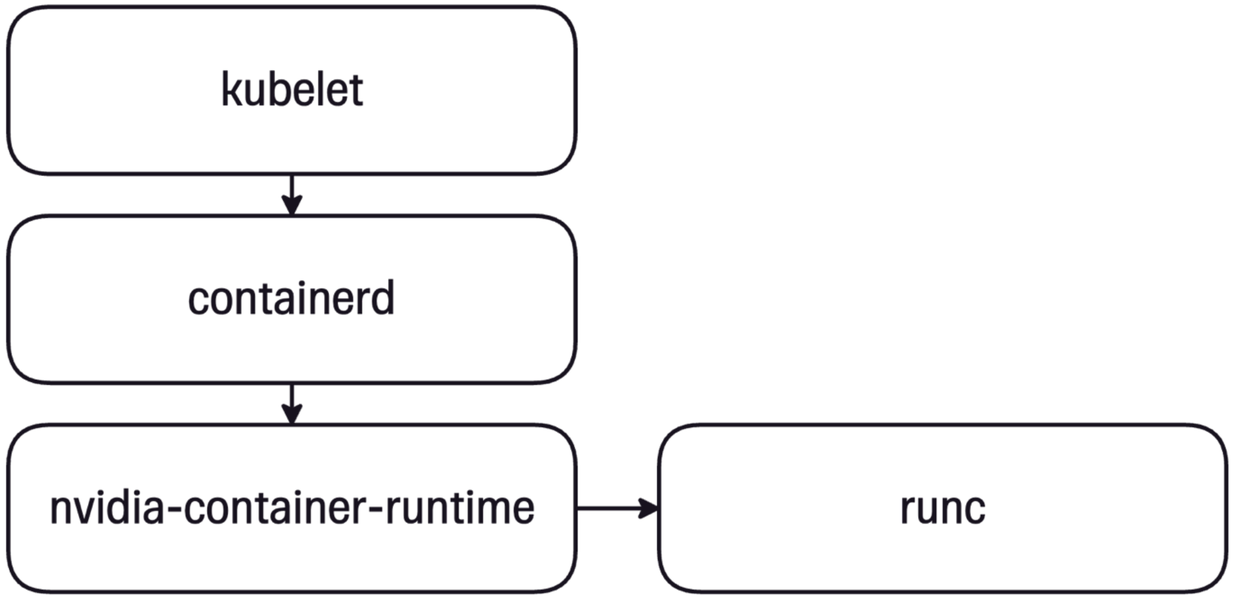
NVIDIA Container Runtime



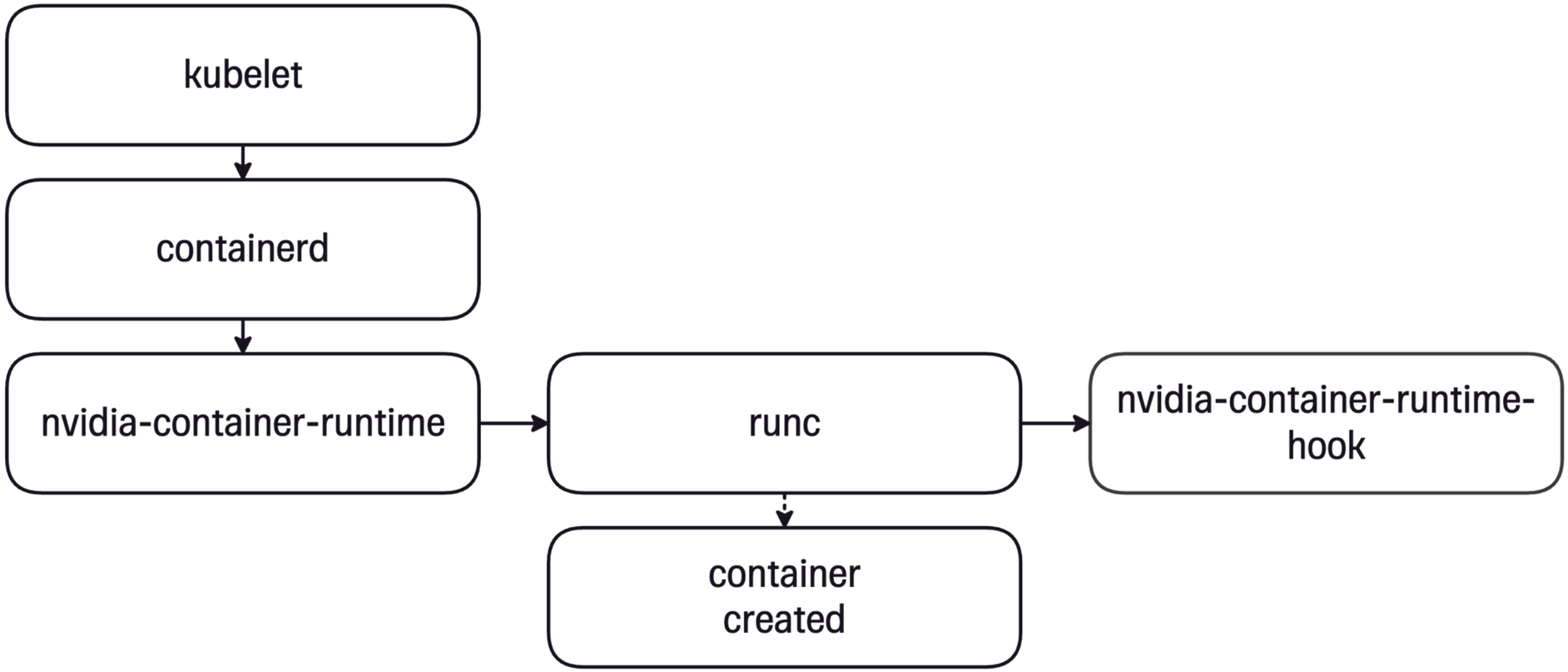
NVIDIA Container Runtime



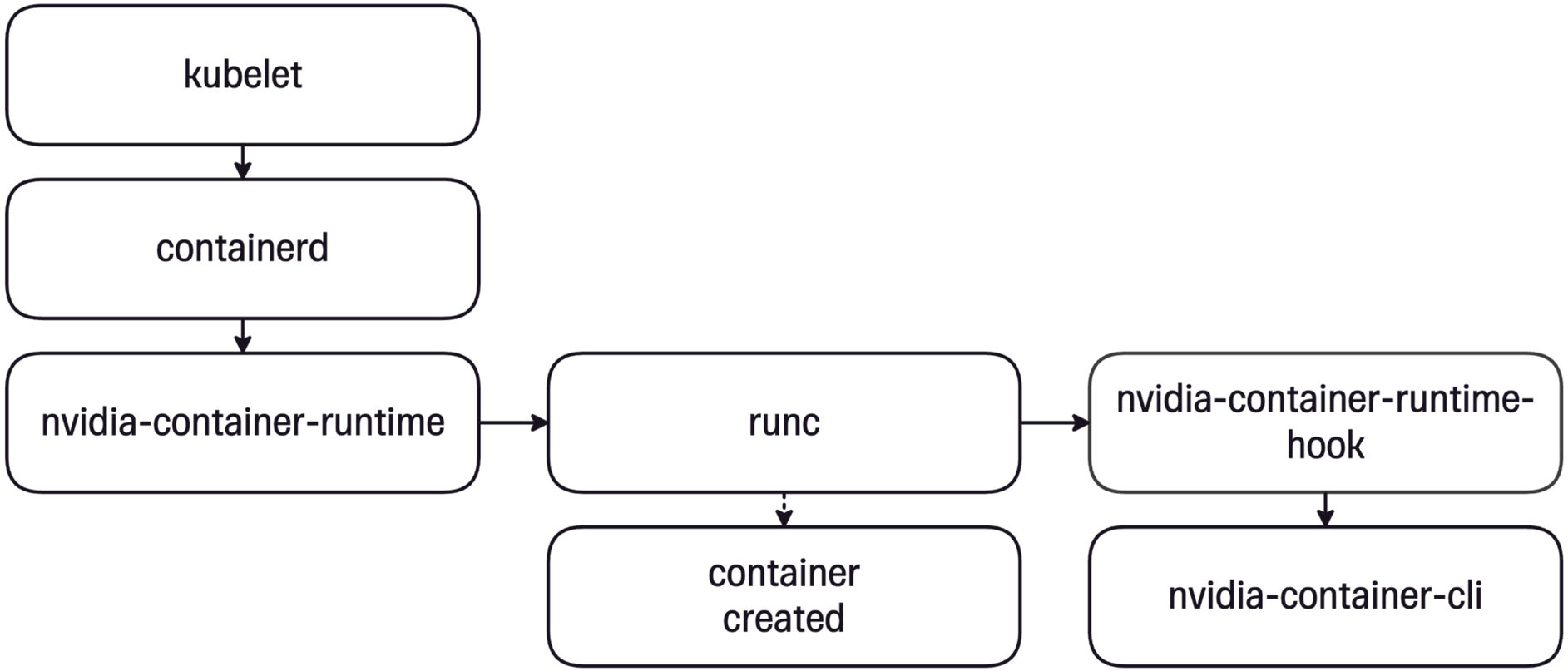
NVIDIA Container Runtime



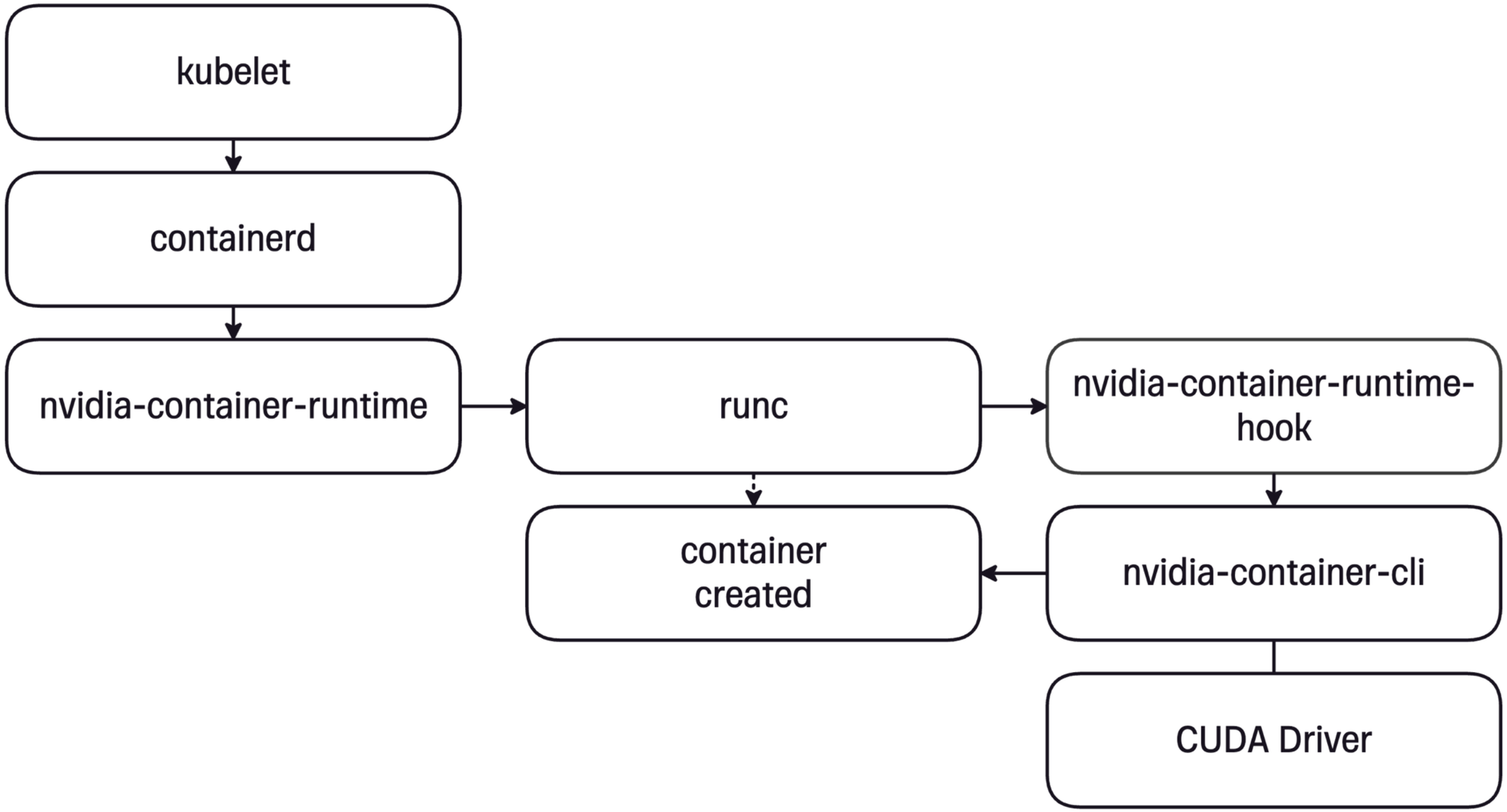
NVIDIA Container Runtime



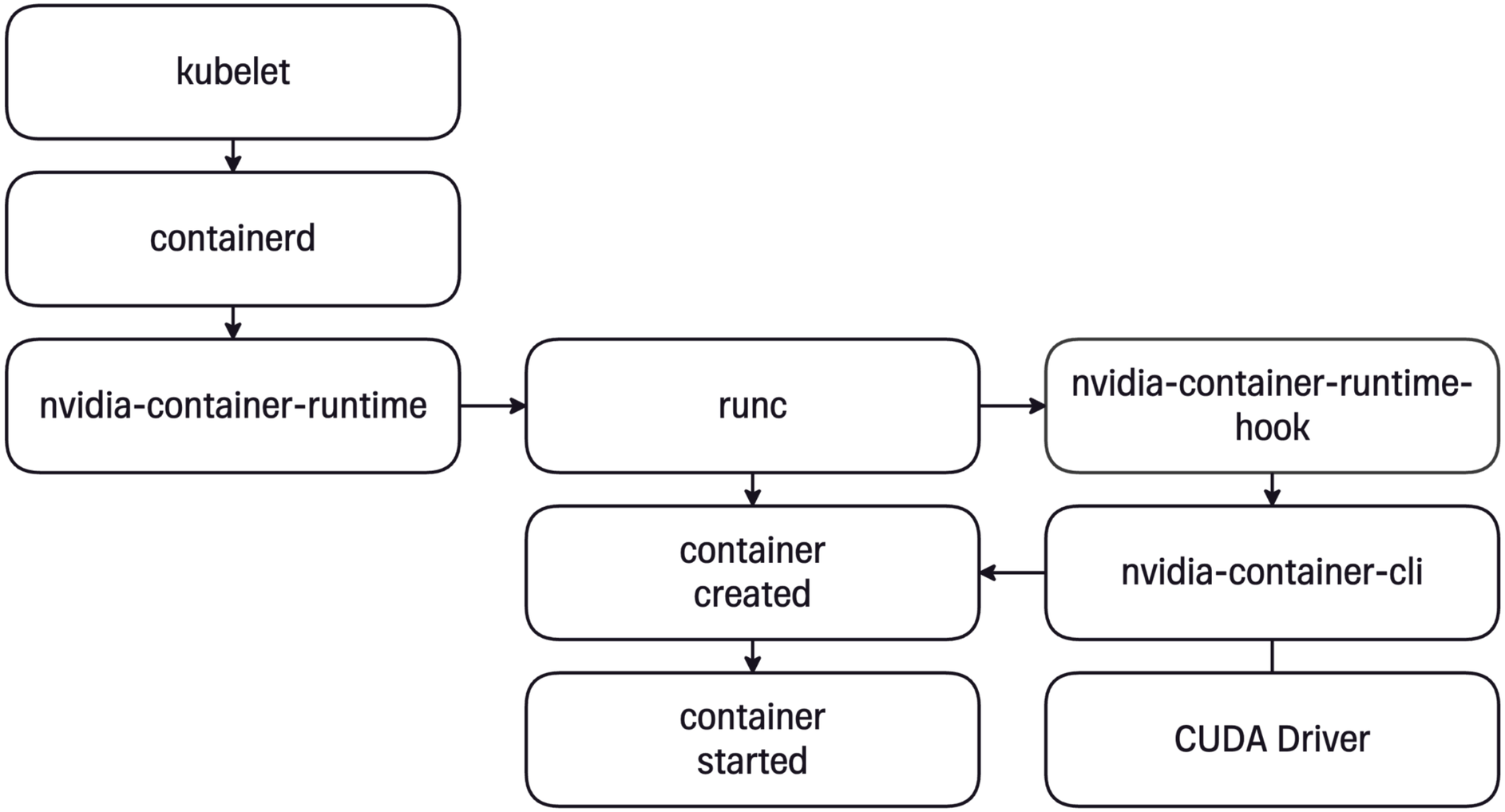
NVIDIA Container Runtime



NVIDIA Container Runtime

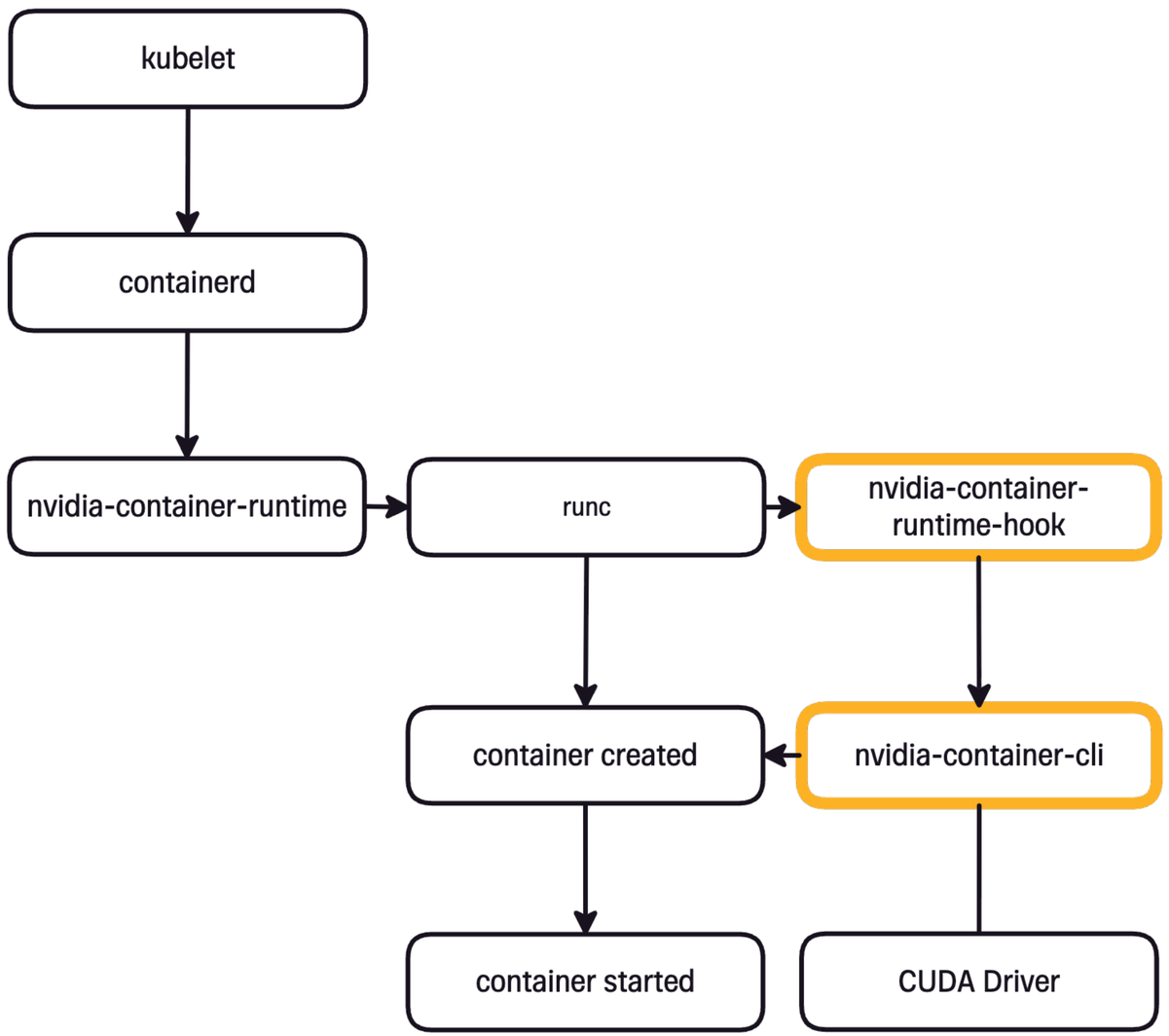


NVIDIA Container Runtime



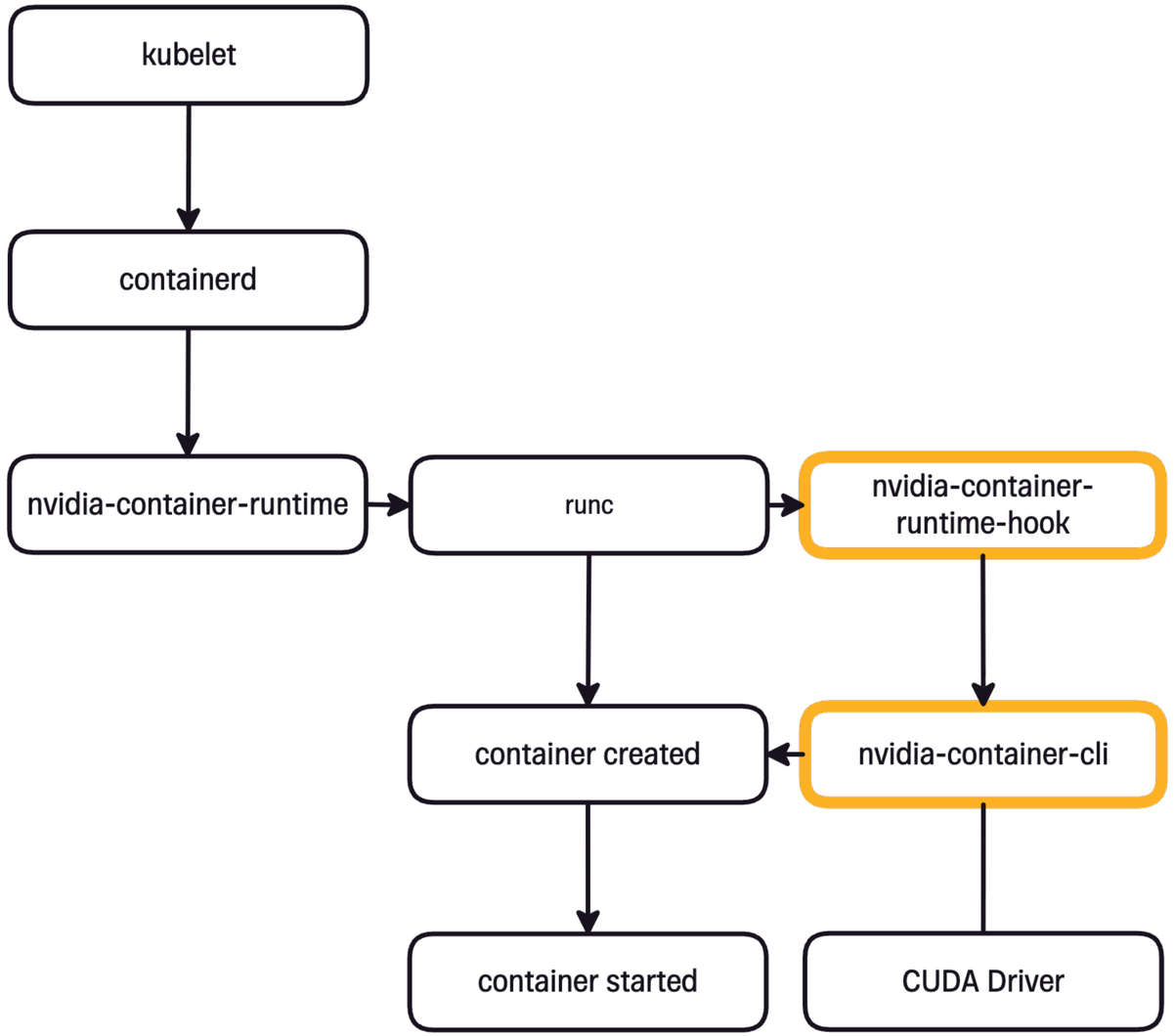
NVIDIA container runtime hook

- Пробрасывает драйвер и CUDA библиотеки



NVIDIA container runtime hook

- Пробрасывает драйвер и CUDA библиотеки
- Настраивает cgroups для доступа к GPU



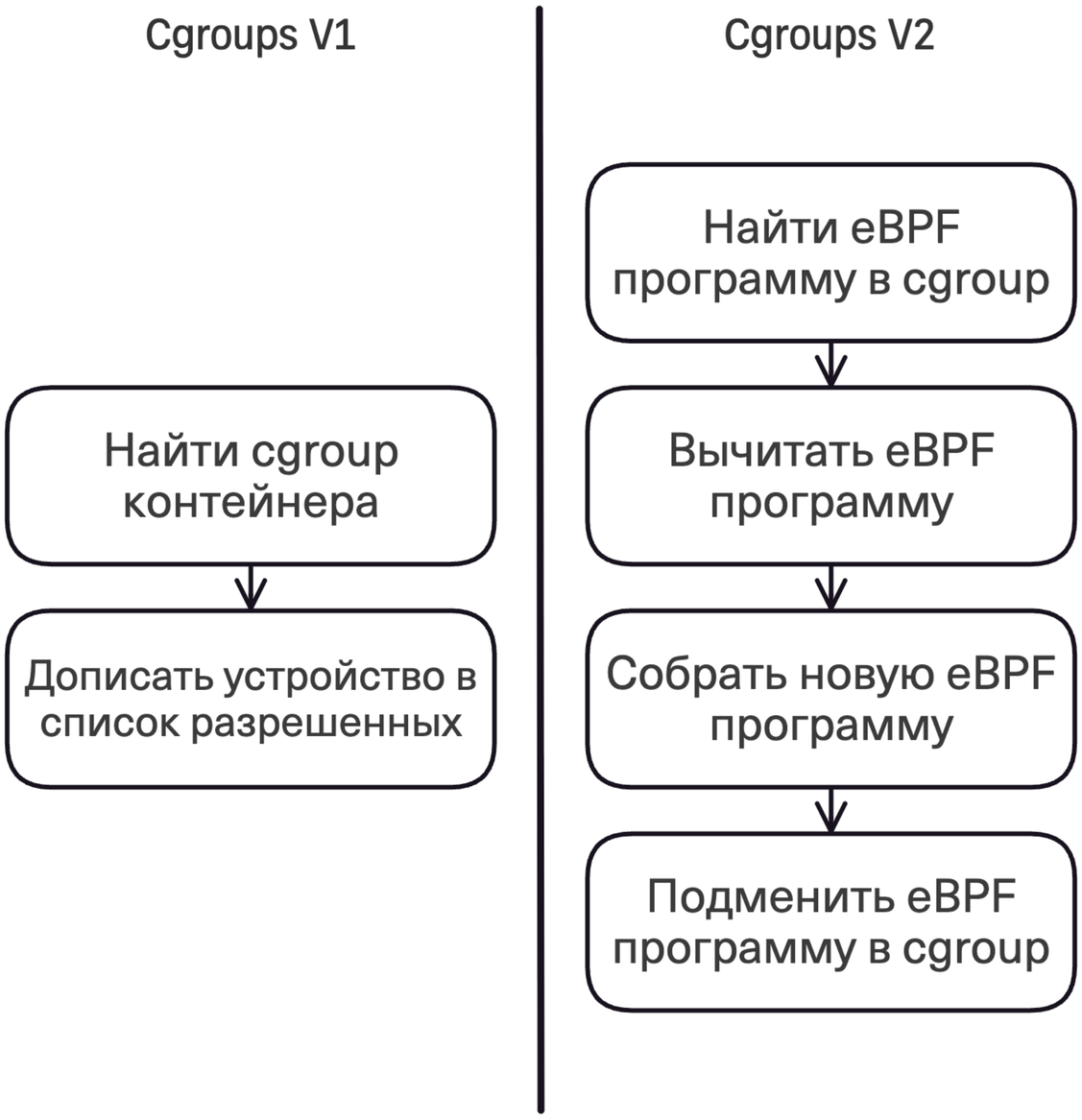
Cgroups для GPU

- Cgroupsv1 – модификация devices.allow



Cgroups для GPU

- Cgroupsv1 – модификация devices.allow
- Cgroupsv2 – генерация eBPF программ и добавление к cgroup



Лечение симптома

Откат на cgroups v1

VS

- Предыдущий кластер успешно работает на cgroups v1
- cgroupsv1 deprecated в Talos 1.10 и Kubernetes 1.31

Лечение симптома

Откат на cgroups v1

- Предыдущий кластер успешно работает на cgroups v1
- cgroupsv1 deprecated в Talos 1.10 и Kubernetes 1.31

VS

Решение root cause

Убрать динамический provisioning

- Видеокарты – статичный ресурс
- CDI – готовое решение

Container Device Interface

Спецификация для container runtime

Единый стандарт описания подключения устройств к container runtime

Container Device Interface

Спецификация для container runtime

Единый стандарт описания подключения устройств к container runtime

Статическая конфигурация

Набор устройств в Kubernetes статичен. Один раз сгенерировал – используешь весь жизненный цикл ноды.

Container Device Interface



Подключение устройств

Как container runtime должен подключать устройства к контейнеру

Container Device Interface



Подключение устройств

Как container runtime должен подключать устройства к контейнеру



Конфигурирование контейнера

Как runc должен дополнительно настраивать контейнер – mount и env

Container Device Interface



Подключение устройств

Как container runtime должен подключать устройства к контейнеру



Конфигурирование контейнера

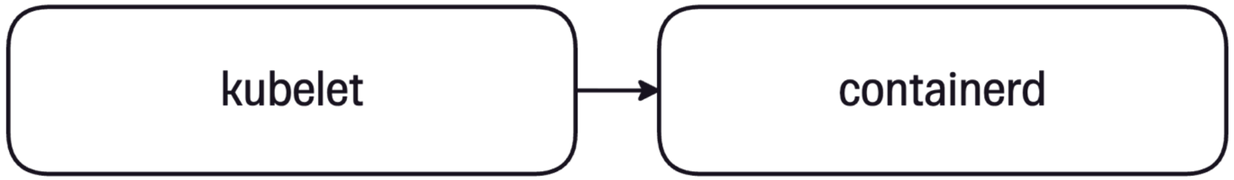
Как runc должен дополнительно настраивать контейнер – mount и env



Легковесные хуки

Дополнительная логика подключения

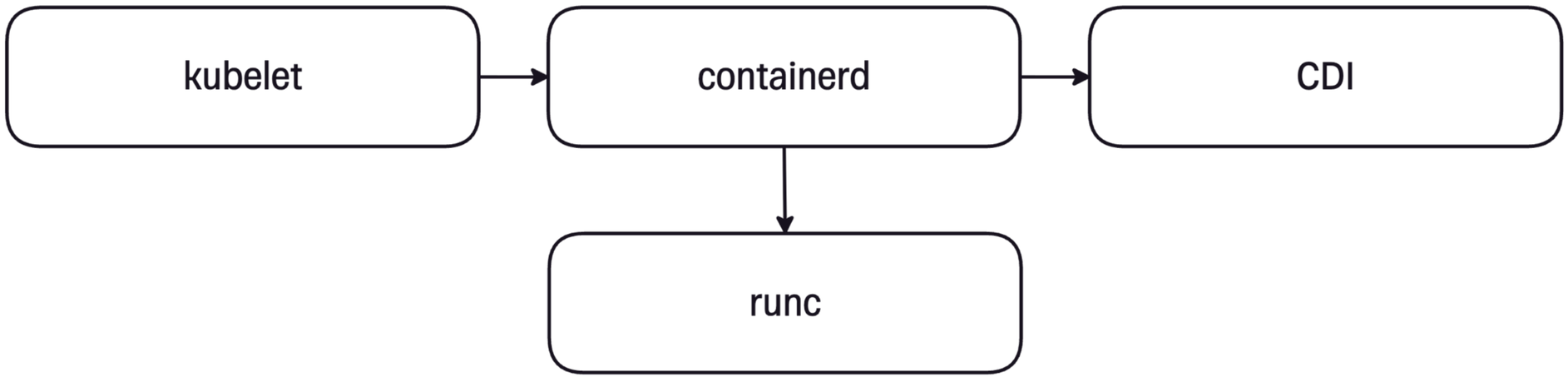
CDI и GPU



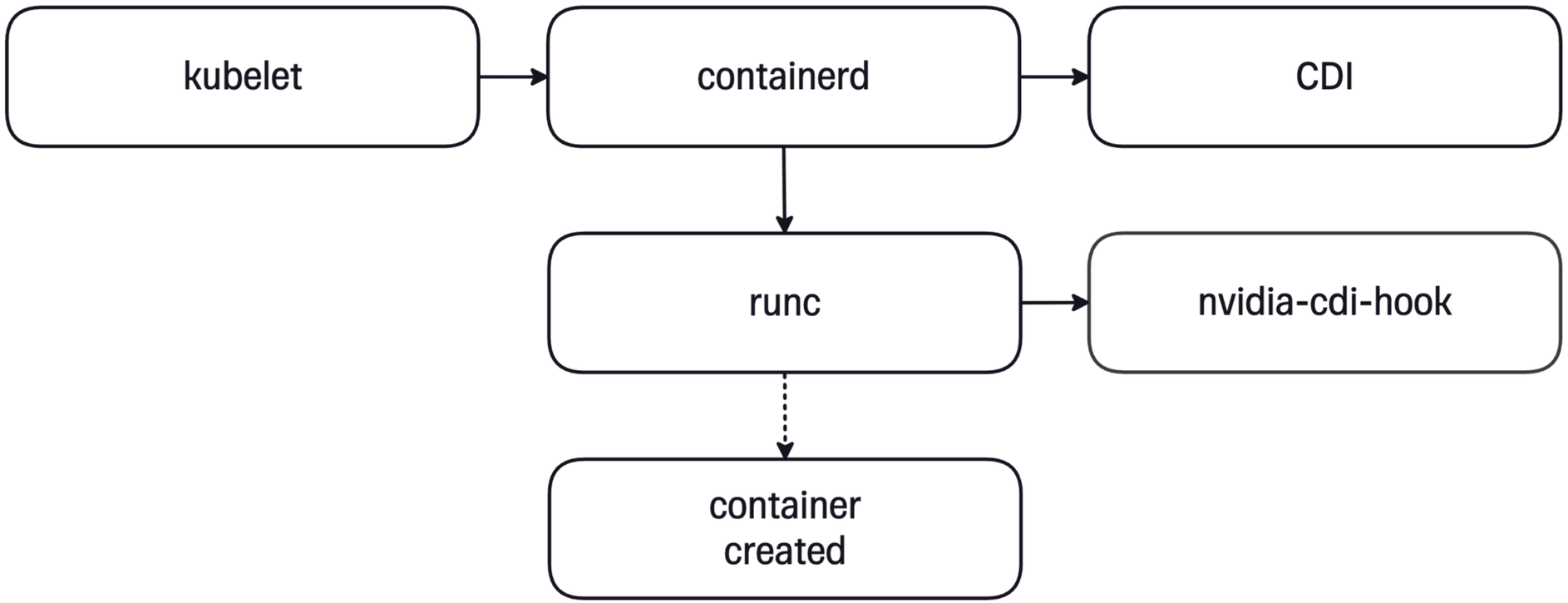
CDI и GPU



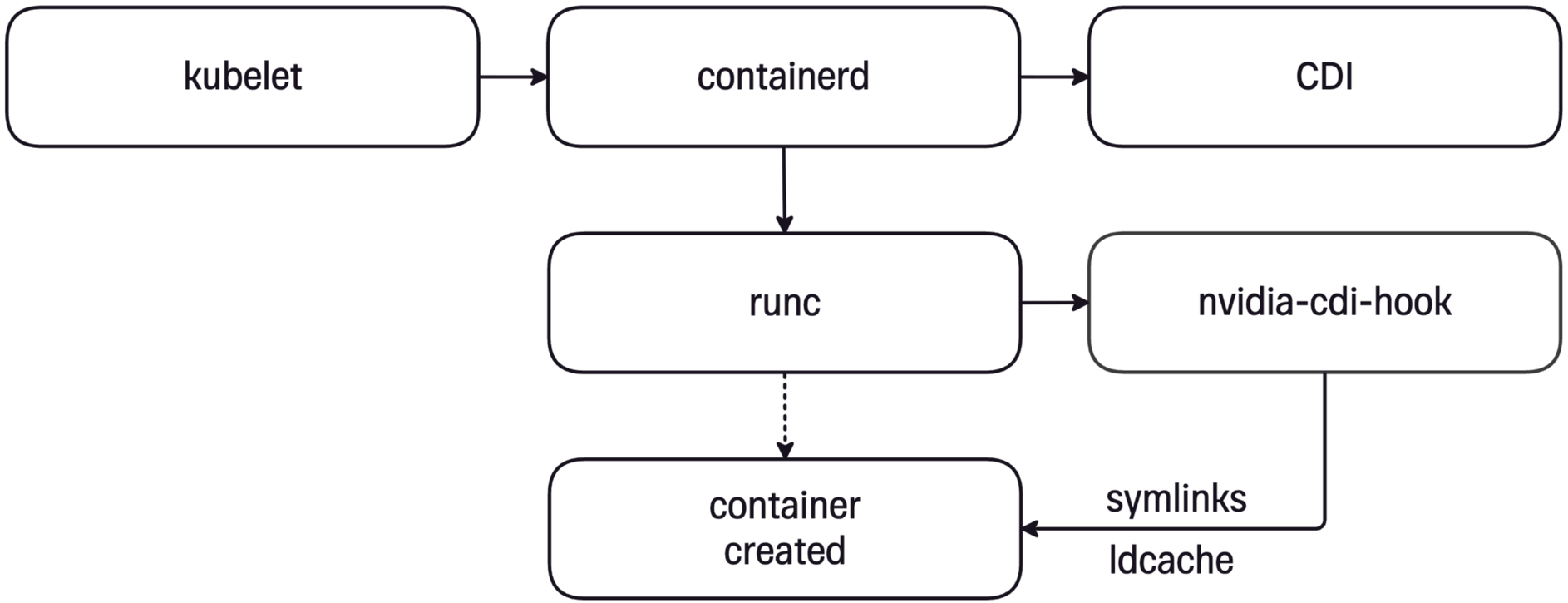
CDI и GPU



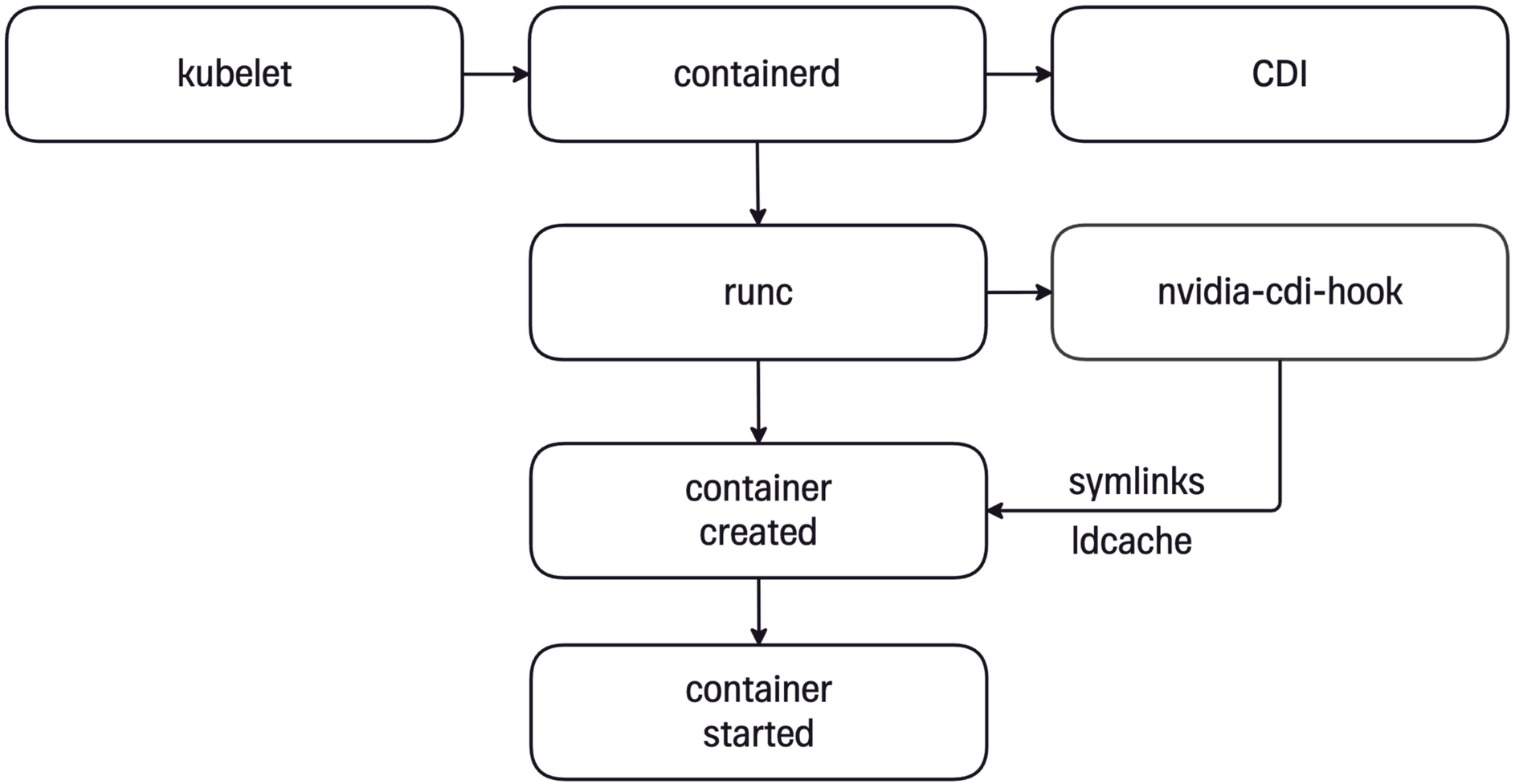
CDI и GPU


kubelet
containerd
CDI
runc
nvidia-cdi-hook
container
created

CDI и GPU

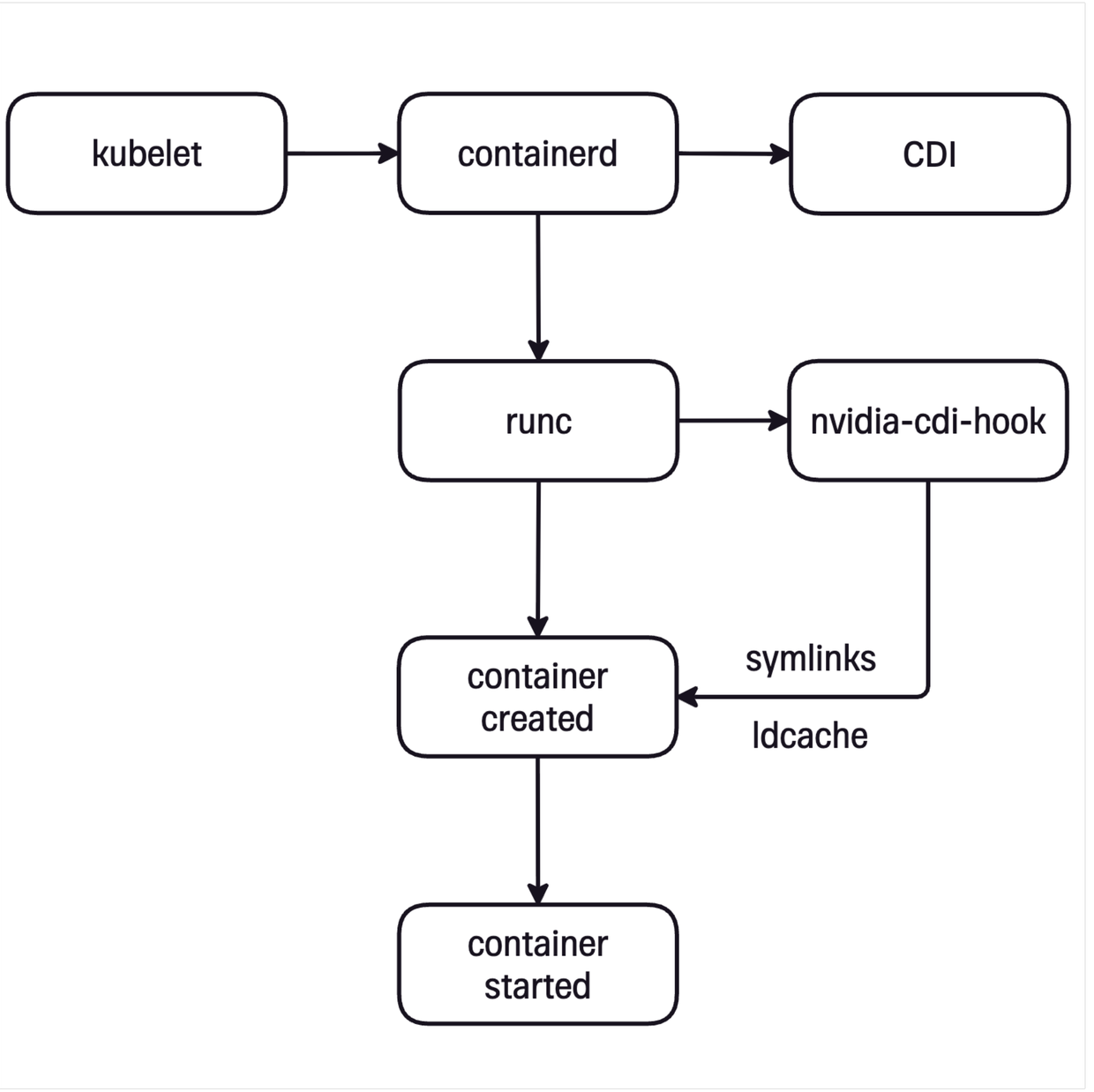

kubelet
containerd
CDI
runc
nvidia-cdi-hook
container
created
symlinks
ldcache

CDI и GPU


kubelet
containerd
CDI
runc
nvidia-cdi-hook
container created
symlinks
ldcache
container started

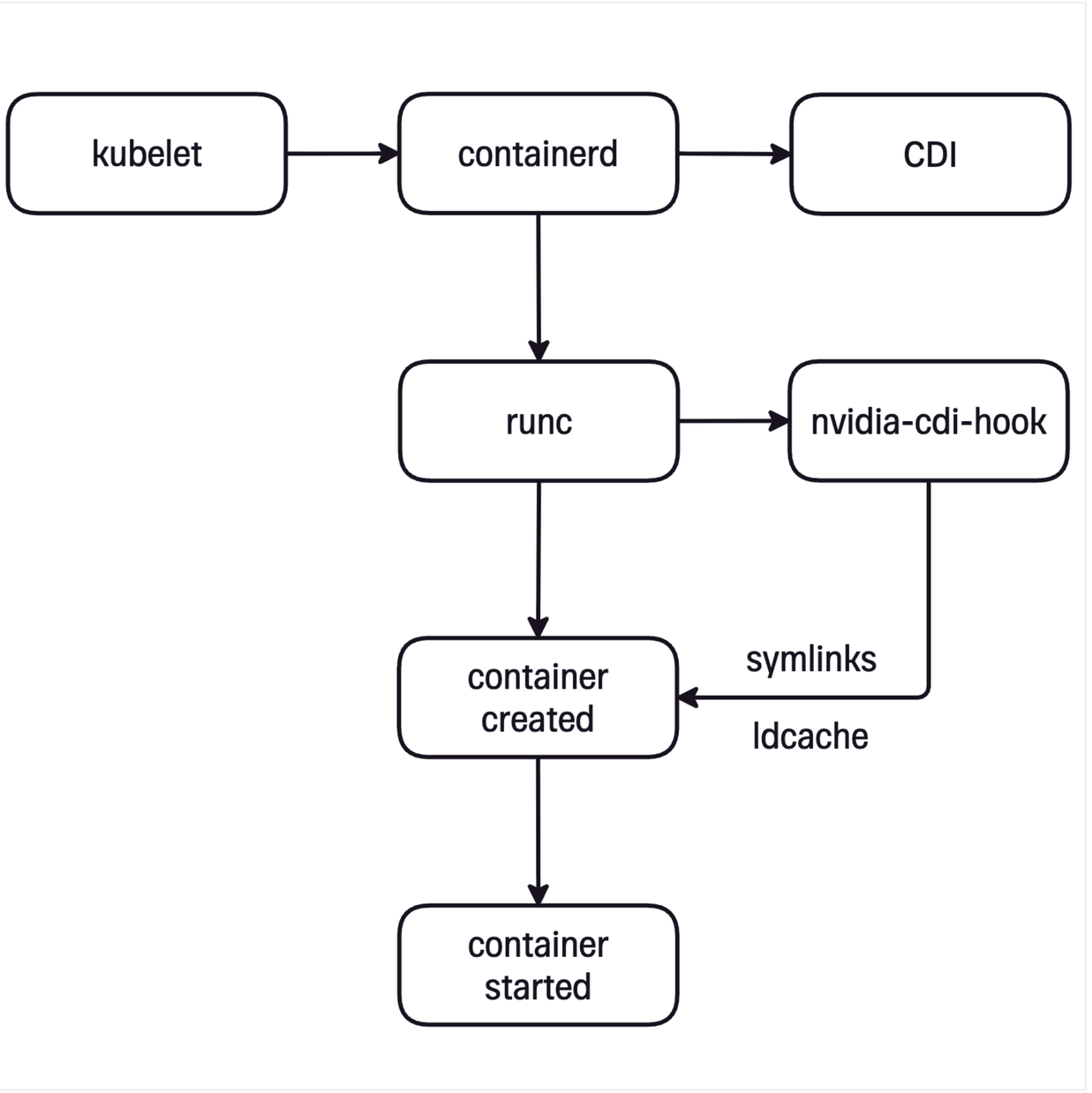
CDI и GPU

- Выполняется статический provisioning



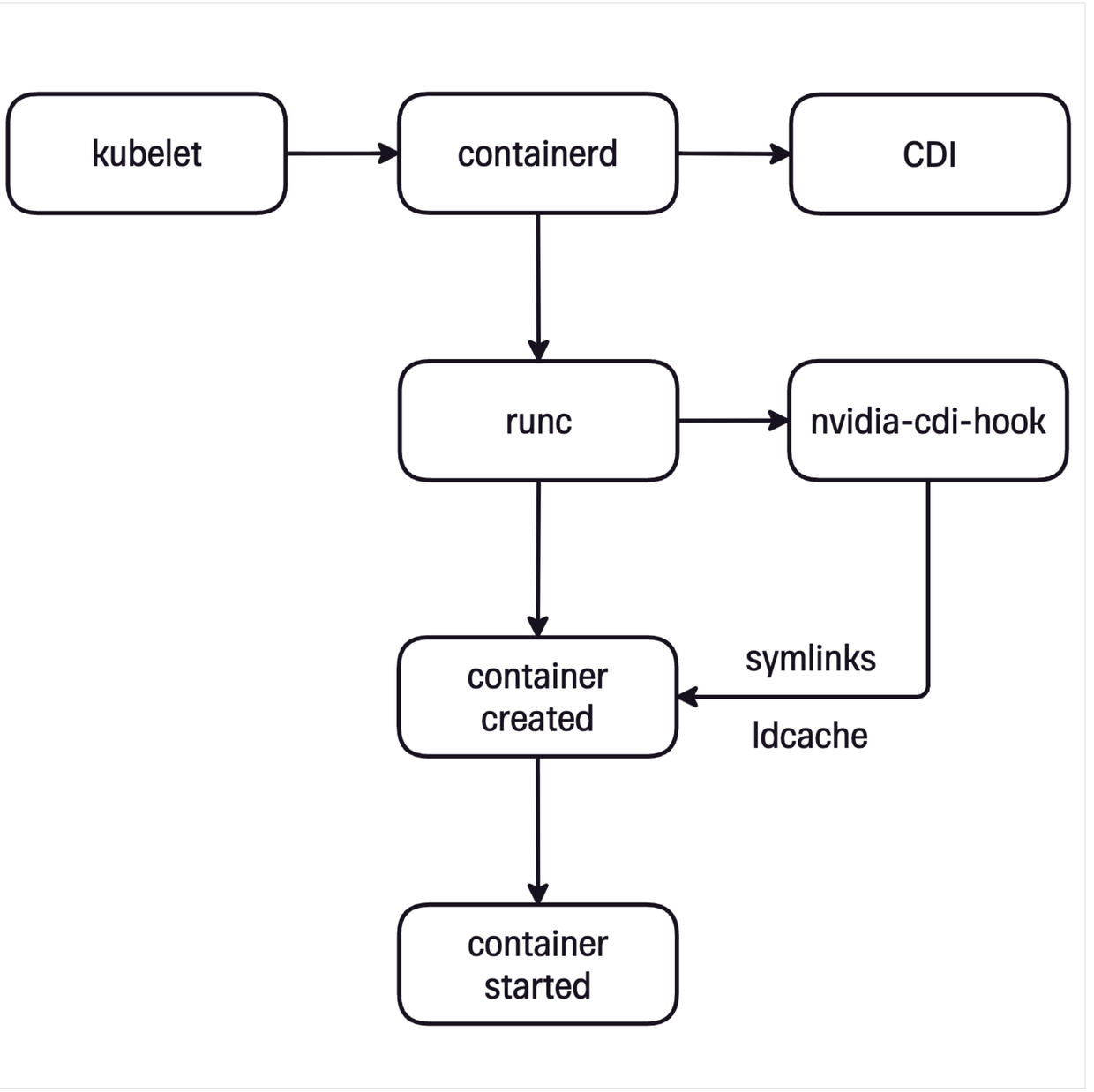
CDI и GPU

- Выполняется статический provisioning
- Hook не выполняет ресурсозатратных действий

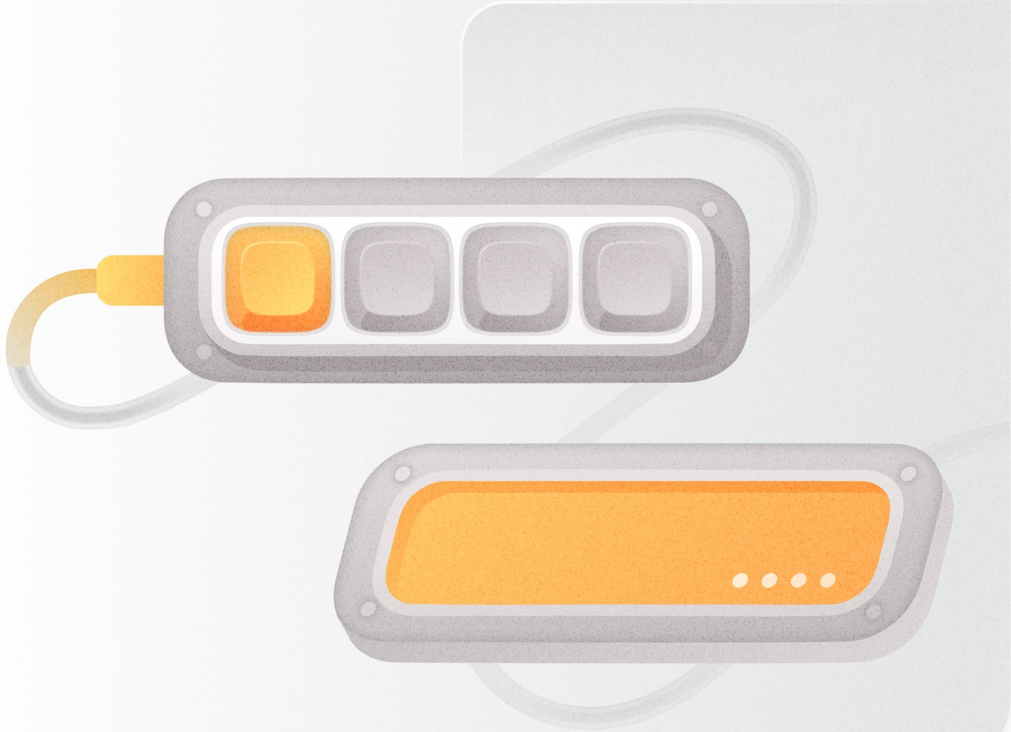


CDI и GPU

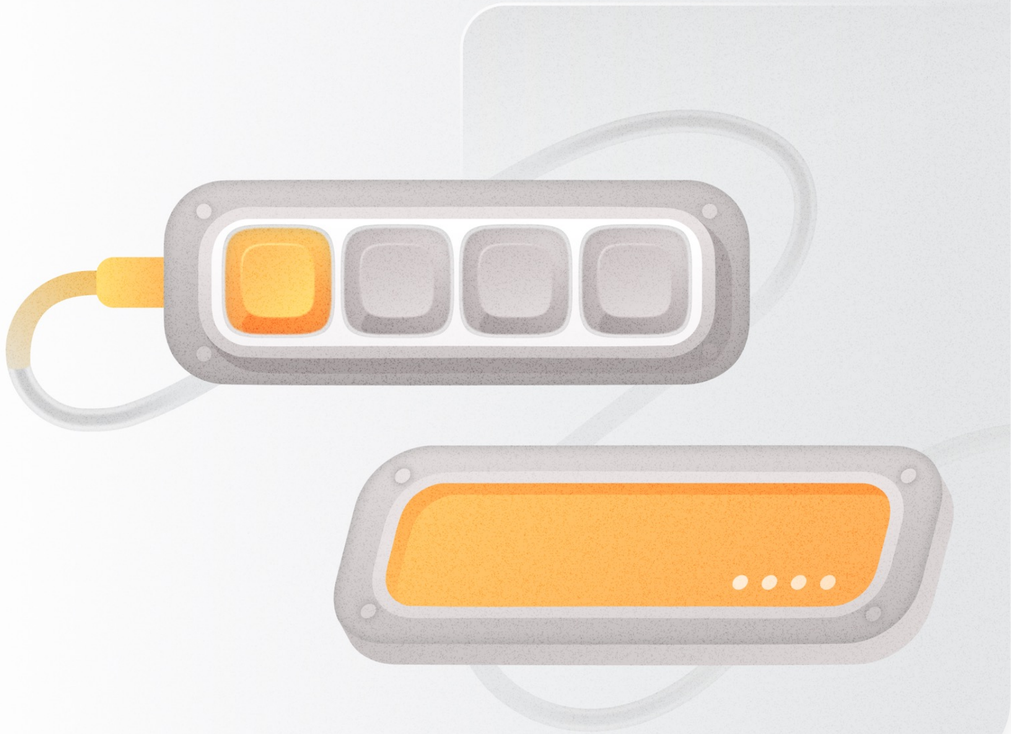
- Выполняется статический provisioning
- Hook не выполняет ресурсозатратных действий
- Cgroups не влияет на работу hook



CDI и GPU Operator



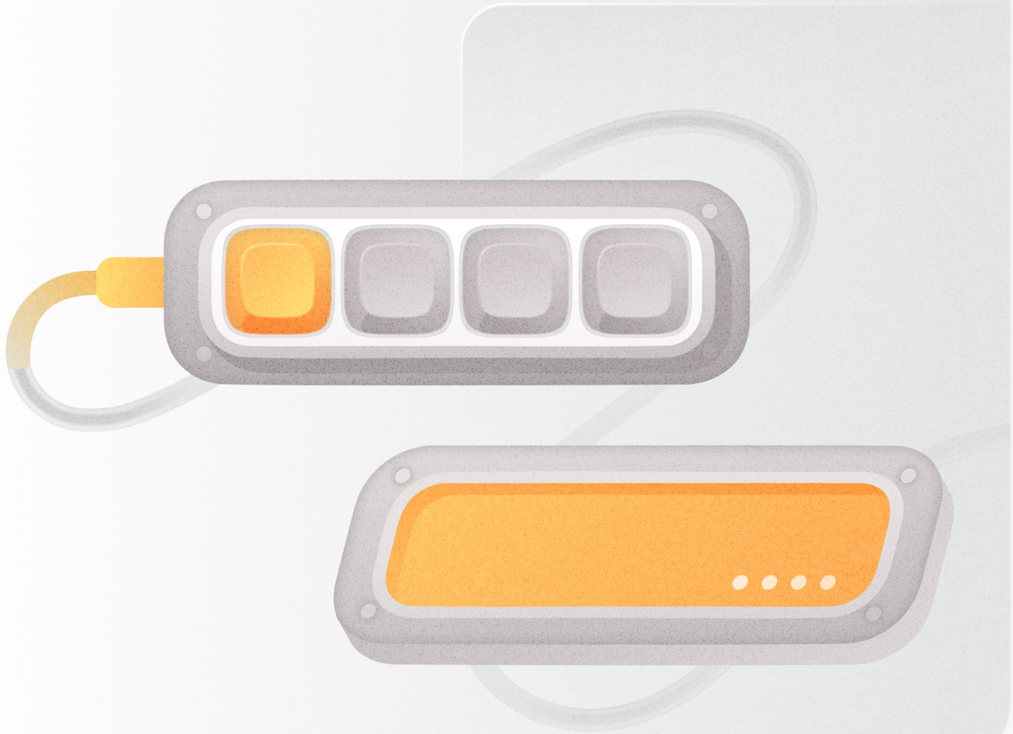
CDI и GPU Operator



CDI by default

В последних версиях GPU Operator CDI используется по умолчанию

CDI и GPU Operator



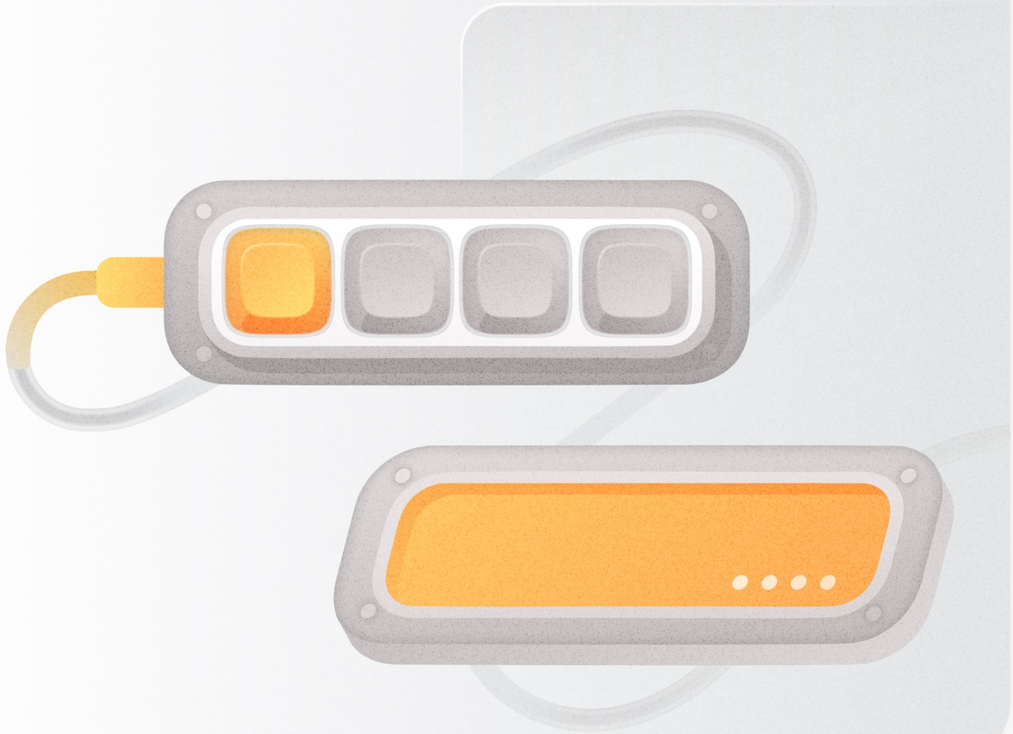
CDI by default

В последних версиях GPU Operator CDI используется по умолчанию

Генерация спецификации

- Container Toolkit – management доступ
- Device Plugin – пользовательский доступ

CDI и GPU Operator



CDI by default

В последних версиях GPU Operator CDI используется по умолчанию

Генерация спецификации

- Container Toolkit – management доступ
- Device Plugin – пользовательский доступ

MIG – новые устройства

После разметки GPU Operator перезапускает генераторы спецификации

CDI и Talos

Несовместимость компонентов

- Device Plugin необходим доступ к драйверу для генерации спецификации

CDI и Talos

Несовместимость компонентов

- Device Plugin необходим доступ к драйверу для генерации спецификации

Нет готового решения

- Issue о поддержке CDI на момент решения проблемы 3 месяца без продвижений
- Поддержку CDI добавили только в Talos 1.13

Запуск GPU Operator: Проблемы

Обнаружение драйвера

- Operator Validator не дает запуститься другим компонентам
- Device Plugin должен сгенерировать CDI спецификацию

Запуск GPU Operator: Проблемы

Обнаружение драйвера

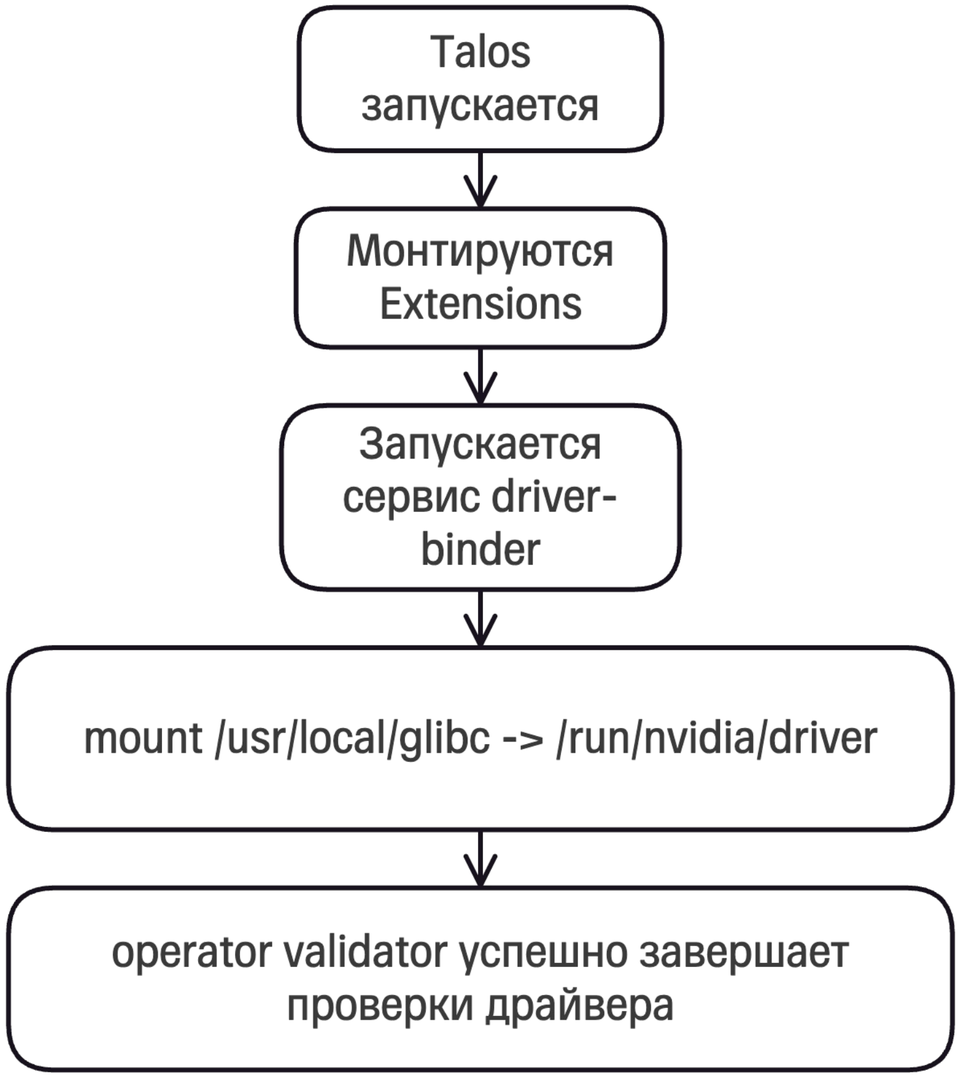
- Operator Validator не дает запуститься другим компонентам
- Device Plugin должен сгенерировать CDI спецификацию

Генерация mgmt CDI спецификации

- Container Toolkit запускается в Legacy режиме
- CDI выключен в Containerd

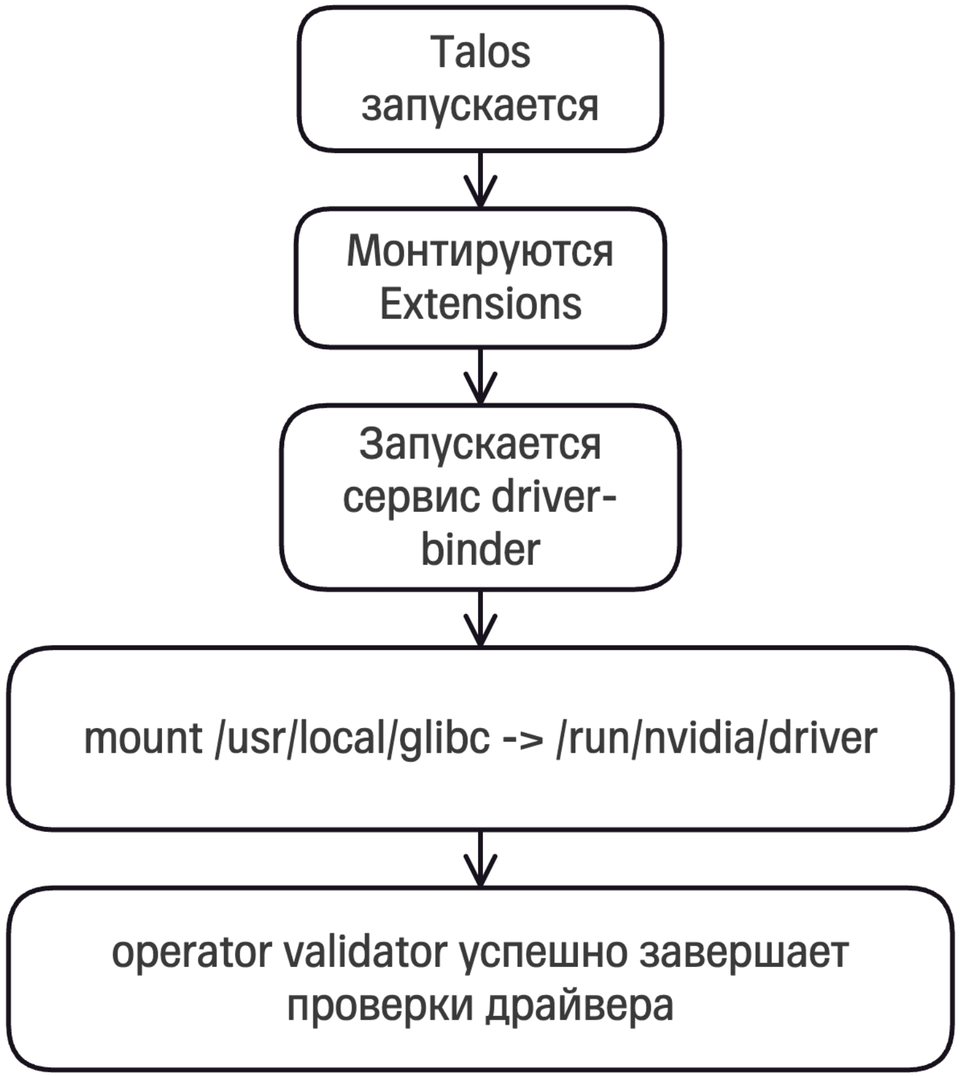
Обнаружение драйвера

- Сервис монтирует библиотеку с драйвером в директорию



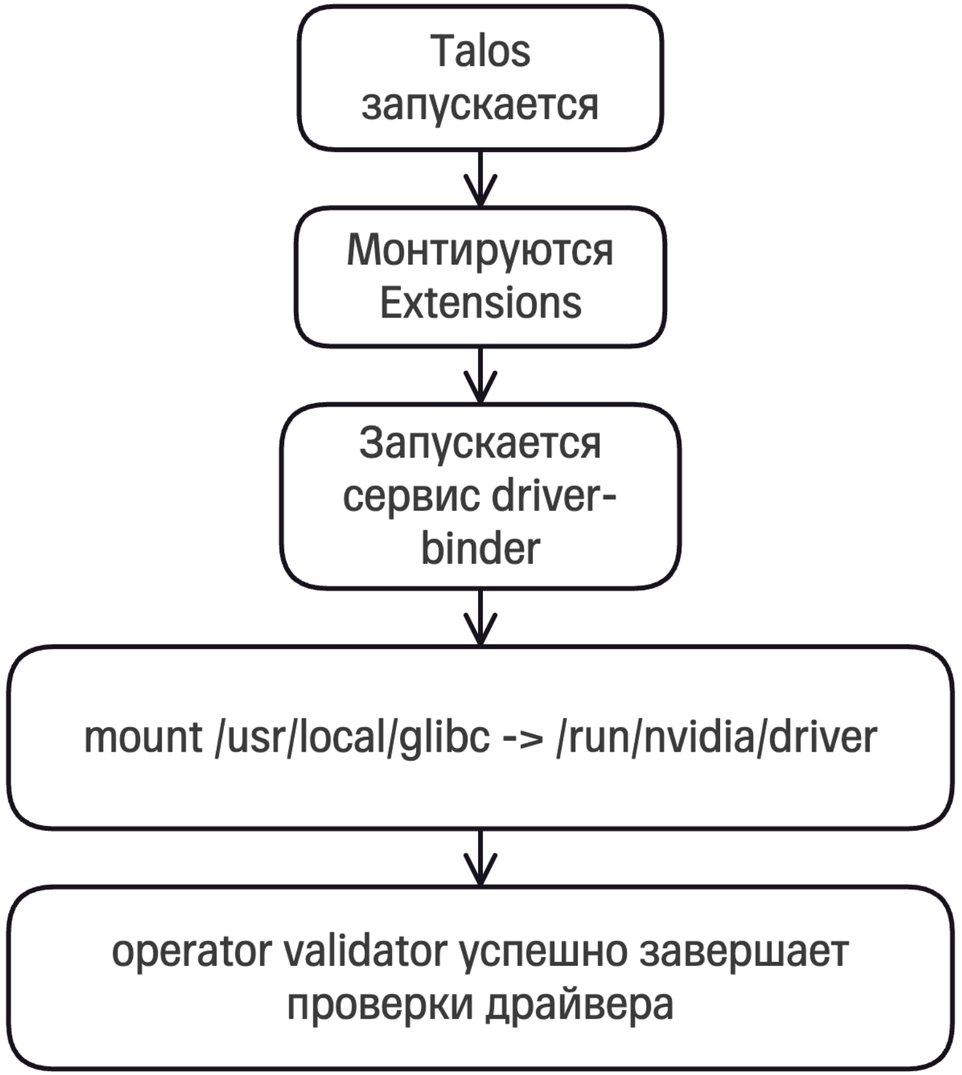
Обнаружение драйвера

- Сервис монтирует библиотеку с драйвером в директорию
- Директория – базовая для установки драйвера средствами GPU Operator



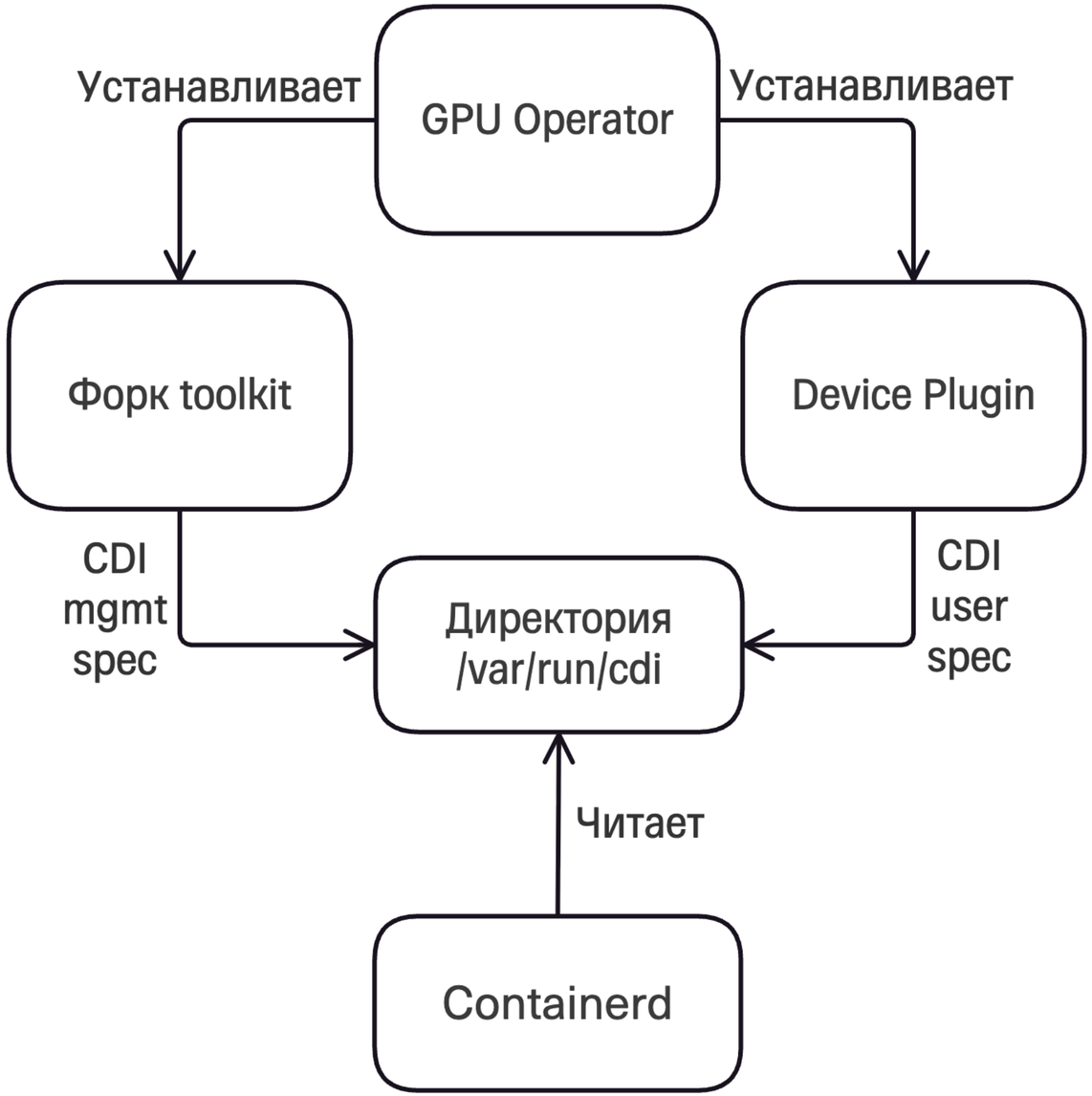
Обнаружение драйвера

- Сервис монтирует библиотеку с драйвером в директорию
- Директория – базовая для установки драйвера средствами GPU Operator
- Operator Validator успешно выполняет проверки



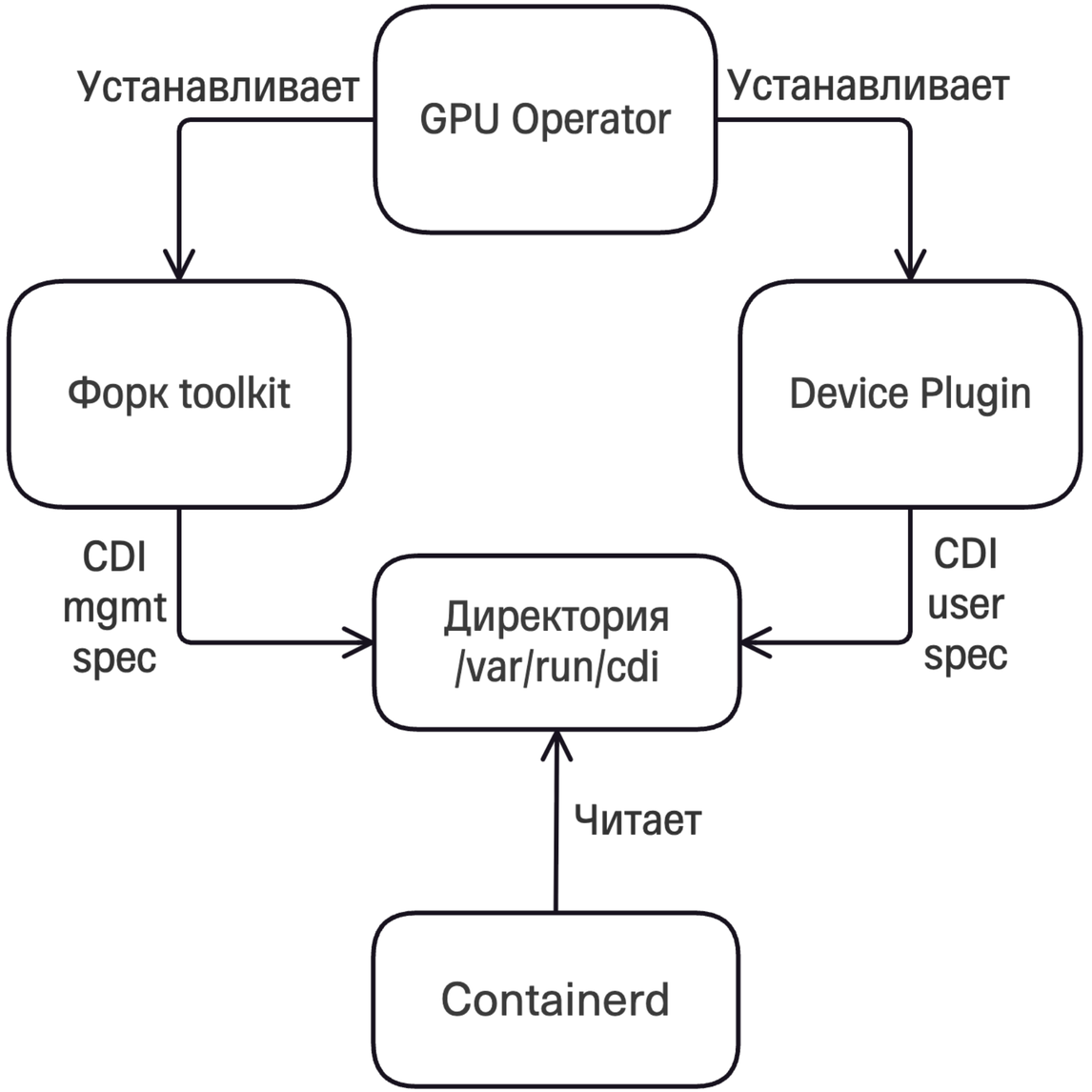
Генерация mgmt CDI спецификации

- Форк toolkit через go обертки позволяет устанавливать его через GPU Operator



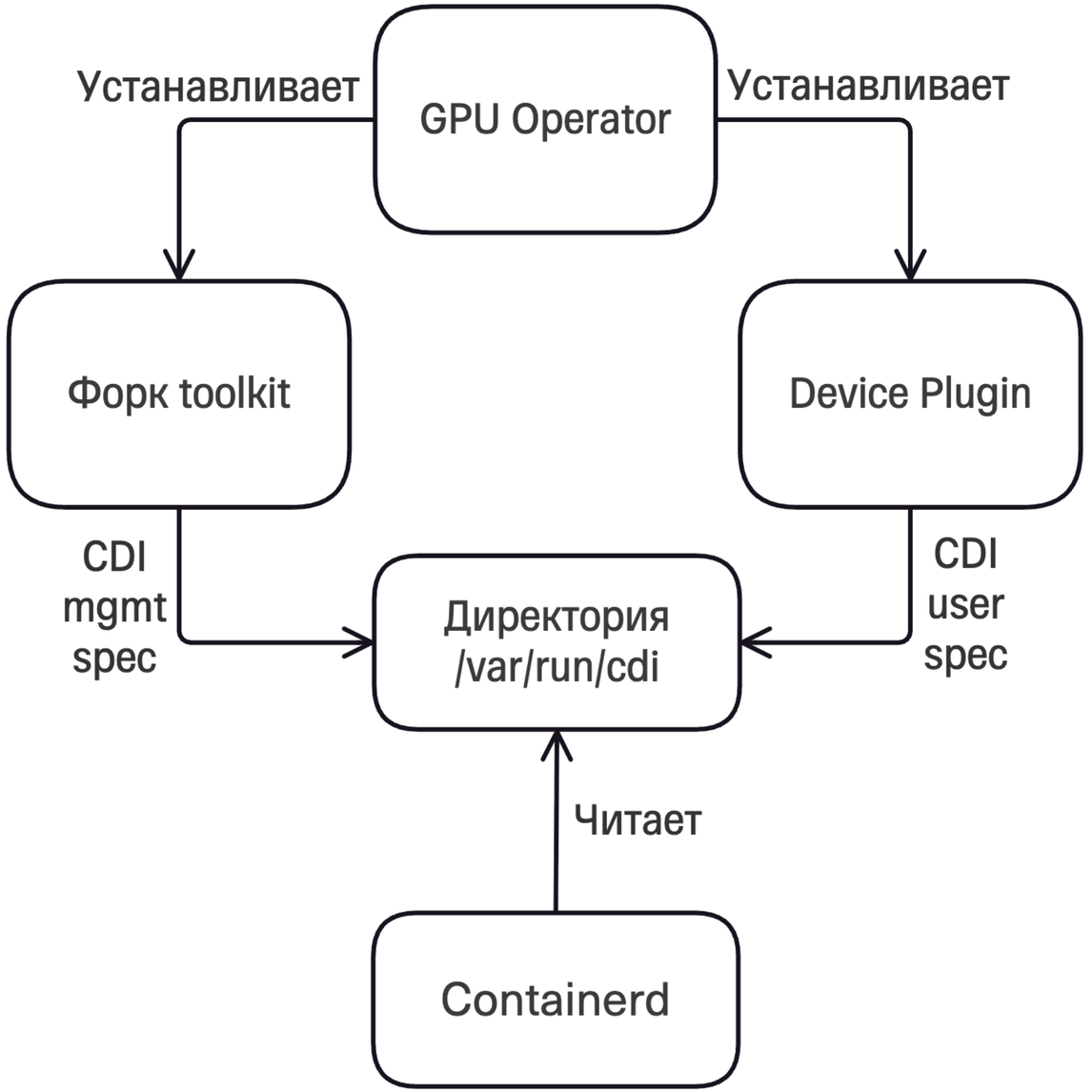
Генерация mgmt CDI спецификации

- Форк toolkit через go обертки позволяет устанавливать его через GPU Operator
- Container Toolkit при запуске генерирует CDI спецификацию

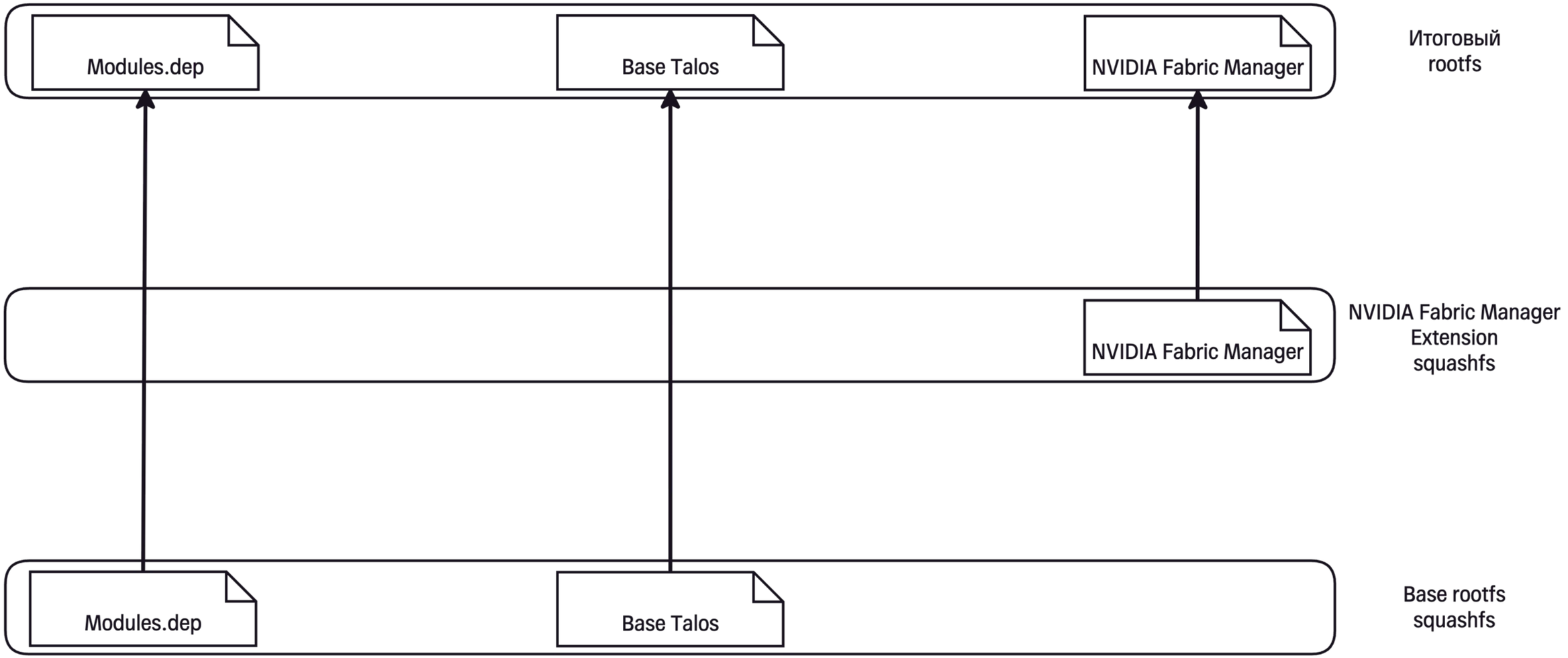


Генерация mgmt CDI спецификации

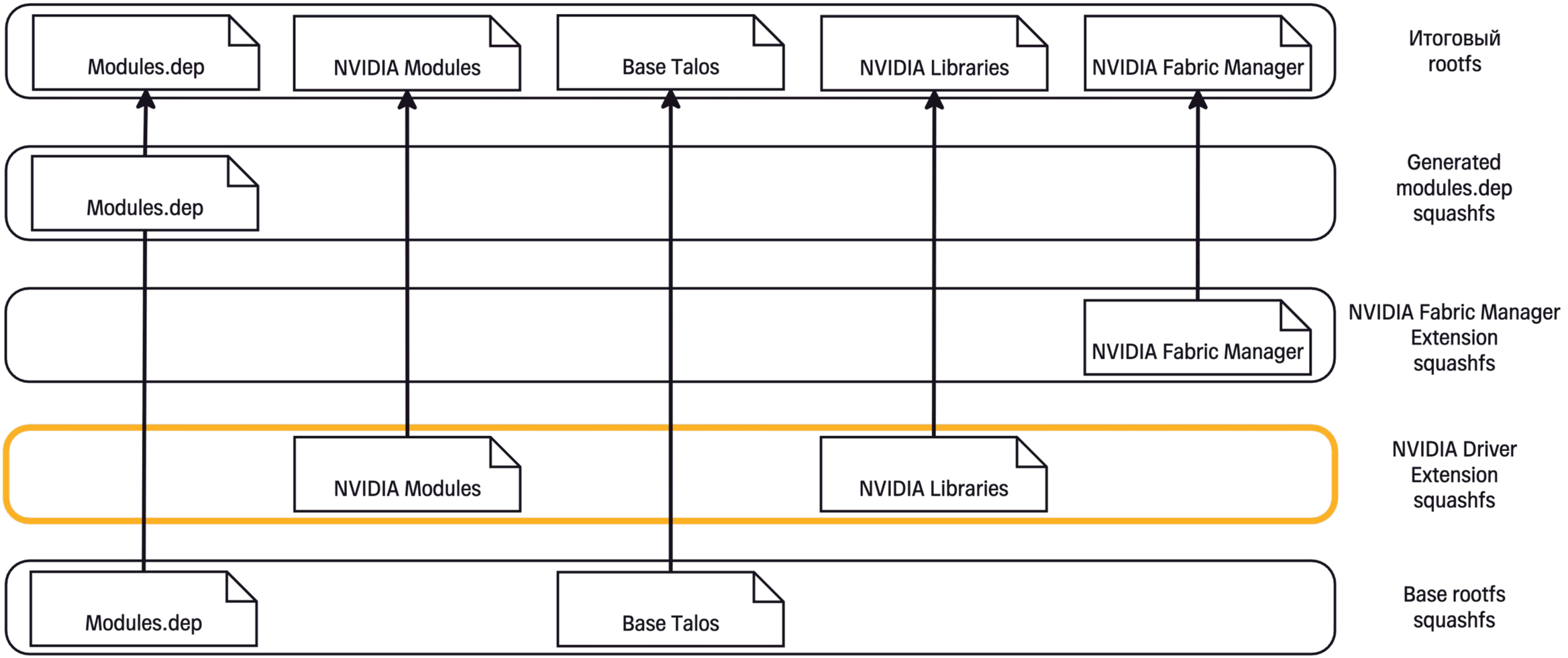
- Форк toolkit через go обертки позволяет устанавливать его через GPU Operator
- Container Toolkit при запуске генерирует CDI спецификацию
- Containerd читает CDI спецификацию из /var/run/cdi – RW директории Talos



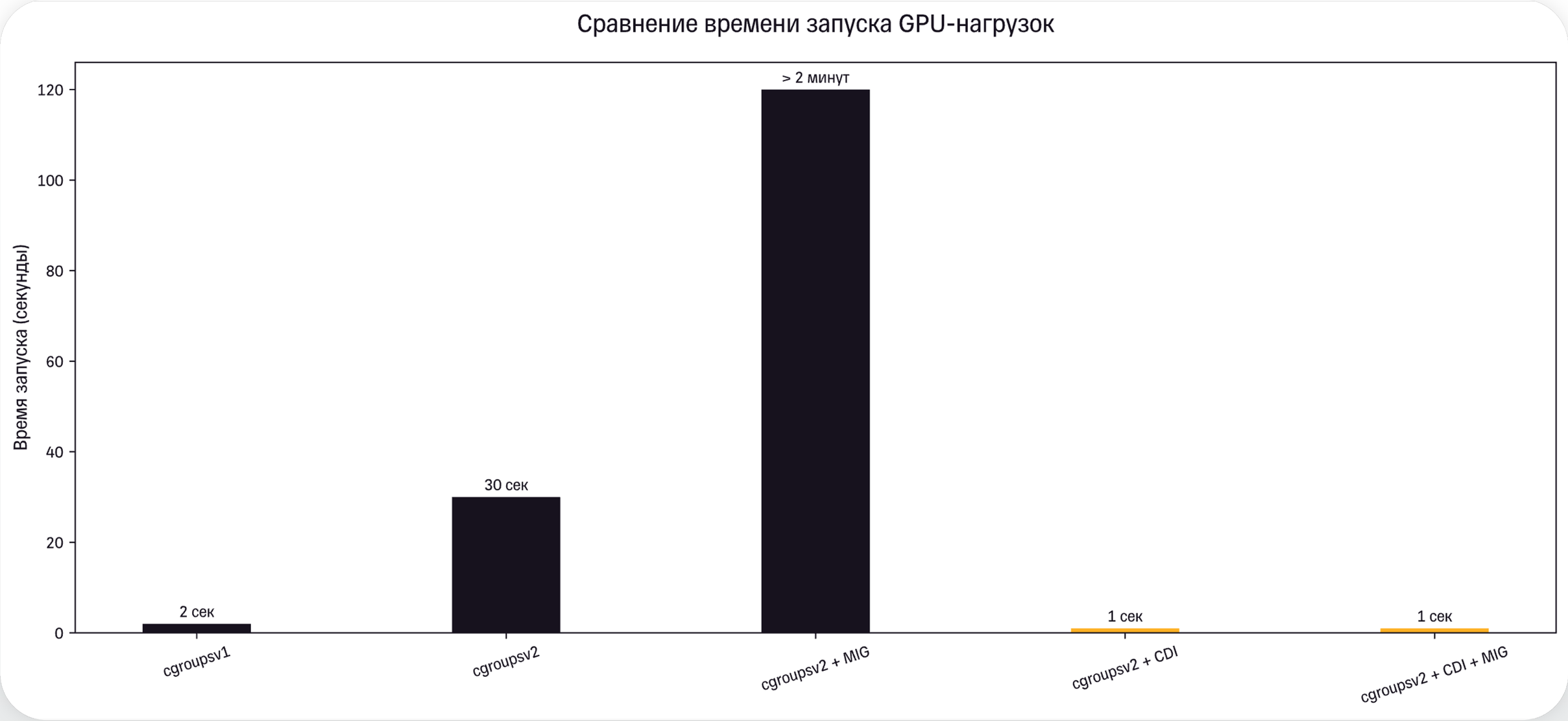
Наше решение



Наше решение



Наше решение





Разработчики предложили другое решение

Наше решение: Выводы

Плюсы

- Для обновления Container Toolkit не нужно обновлять Talos



Минусы

- Большое количество форков, которые не пошли в апстрим



Talos 1.13+

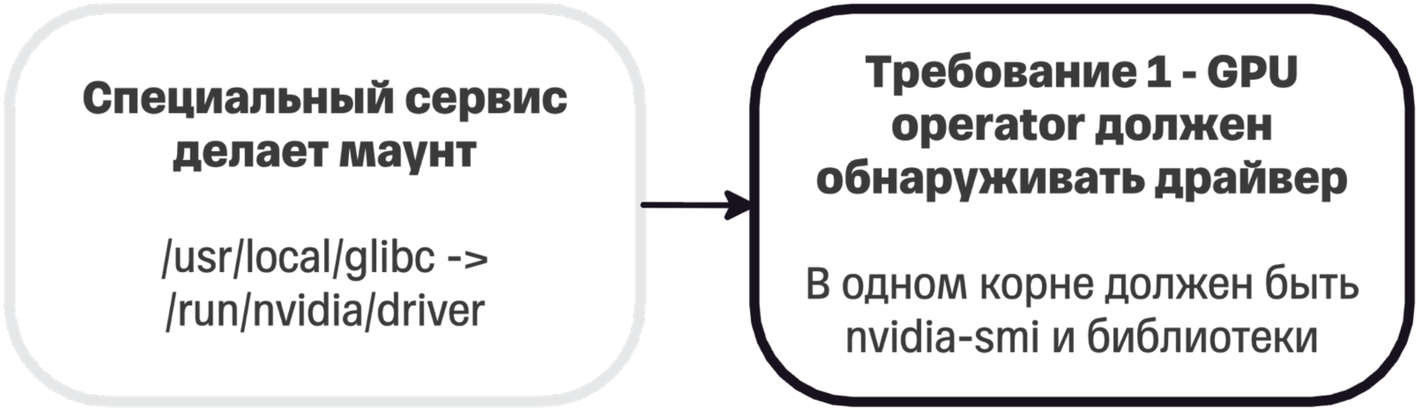
Требование 1 - GPU operator должен обнаруживать драйвер

В одном корне должен быть nvidia-smi и библиотеки

Наше решение

Talos 1.13+

Talos 1.13+



Наше решение

Talos 1.13+

Talos 1.13+



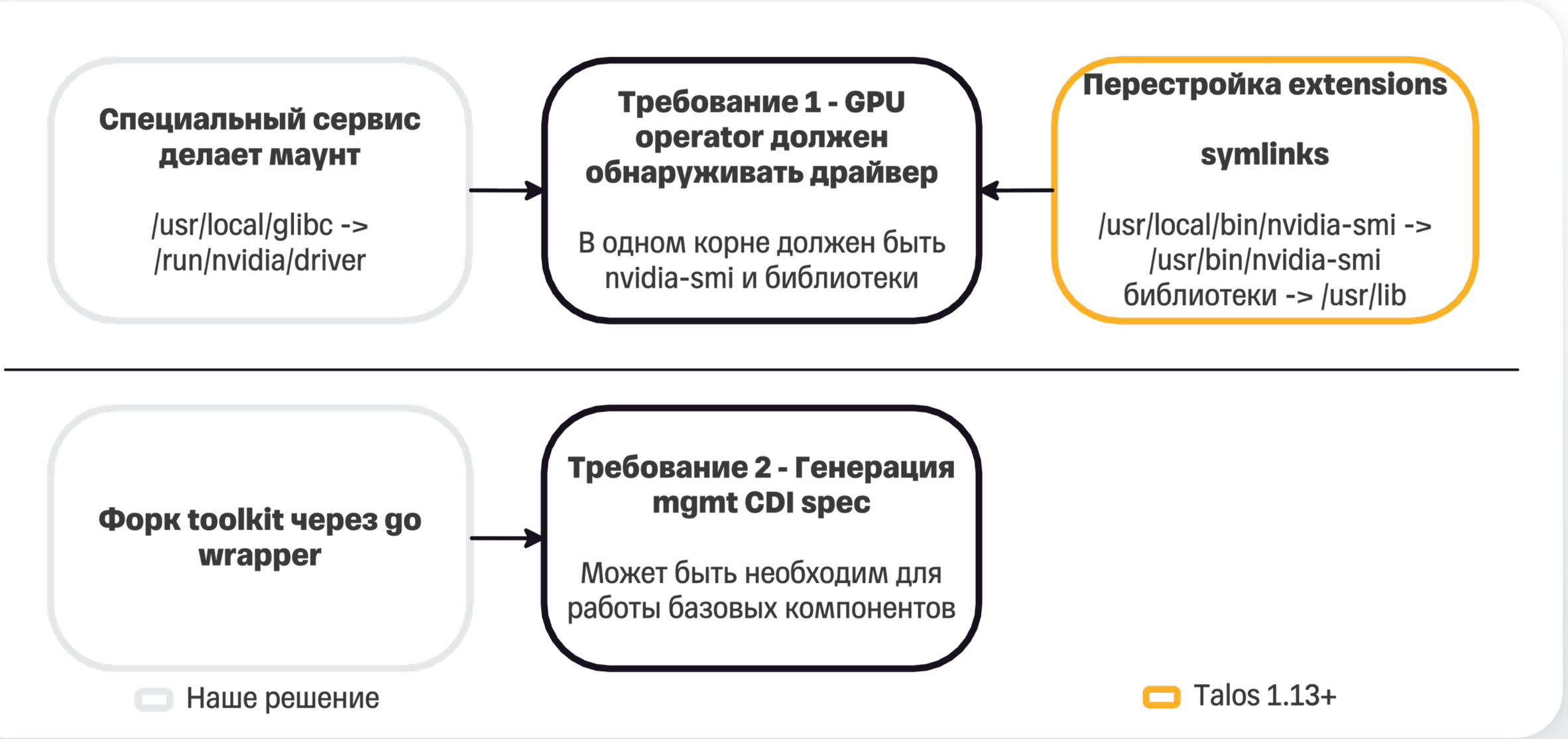
Наше решение

Talos 1.13+

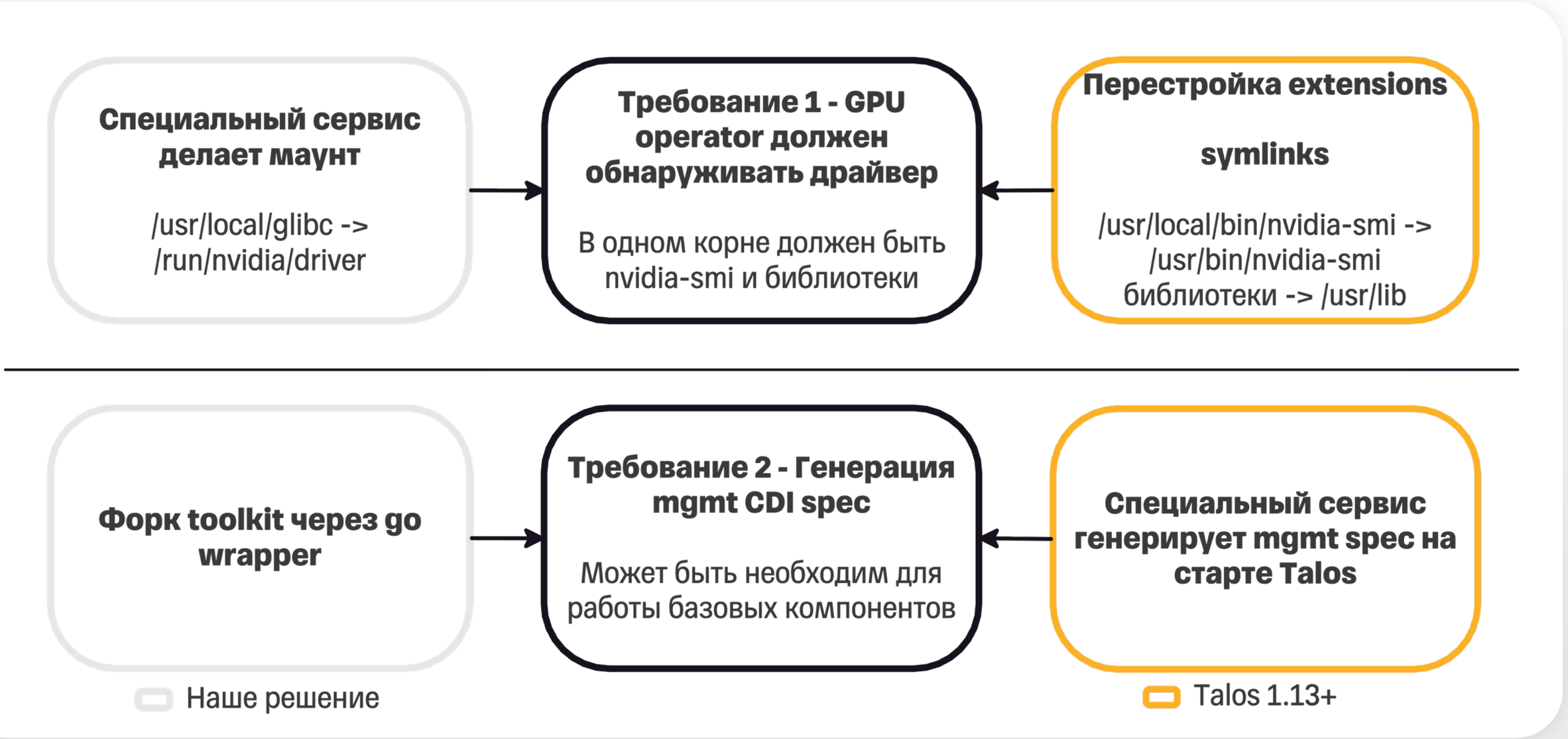
Talos 1.13+


Специальный сервис делает маунт
/usr/local/glibc -> /run/nvidia/driver
Требование 1 - GPU operator должен обнаруживать драйвер
В одном корне должен быть nvidia-smi и библиотеки
Перестройка extensions
symlinks
/usr/local/bin/nvidia-smi -> /usr/bin/nvidia-smi
библиотеки -> /usr/lib
Требование 2 - Генерация mgmt CDI spec
Может быть необходим для работы базовых компонентов
Наше решение
Talos 1.13+

Talos 1.13+


Специальный сервис делает маунт
/usr/local/glibc -> /run/nvidia/driver
Требование 1 - GPU operator должен обнаруживать драйвер
В одном корне должен быть nvidia-smi и библиотеки
Перестройка extensions
symlinks
/usr/local/bin/nvidia-smi -> /usr/bin/nvidia-smi
библиотеки -> /usr/lib
Форк toolkit через go wrapper
Требование 2 - Генерация mgmt CDI spec
Может быть необходим для работы базовых компонентов
Наше решение
Talos 1.13+

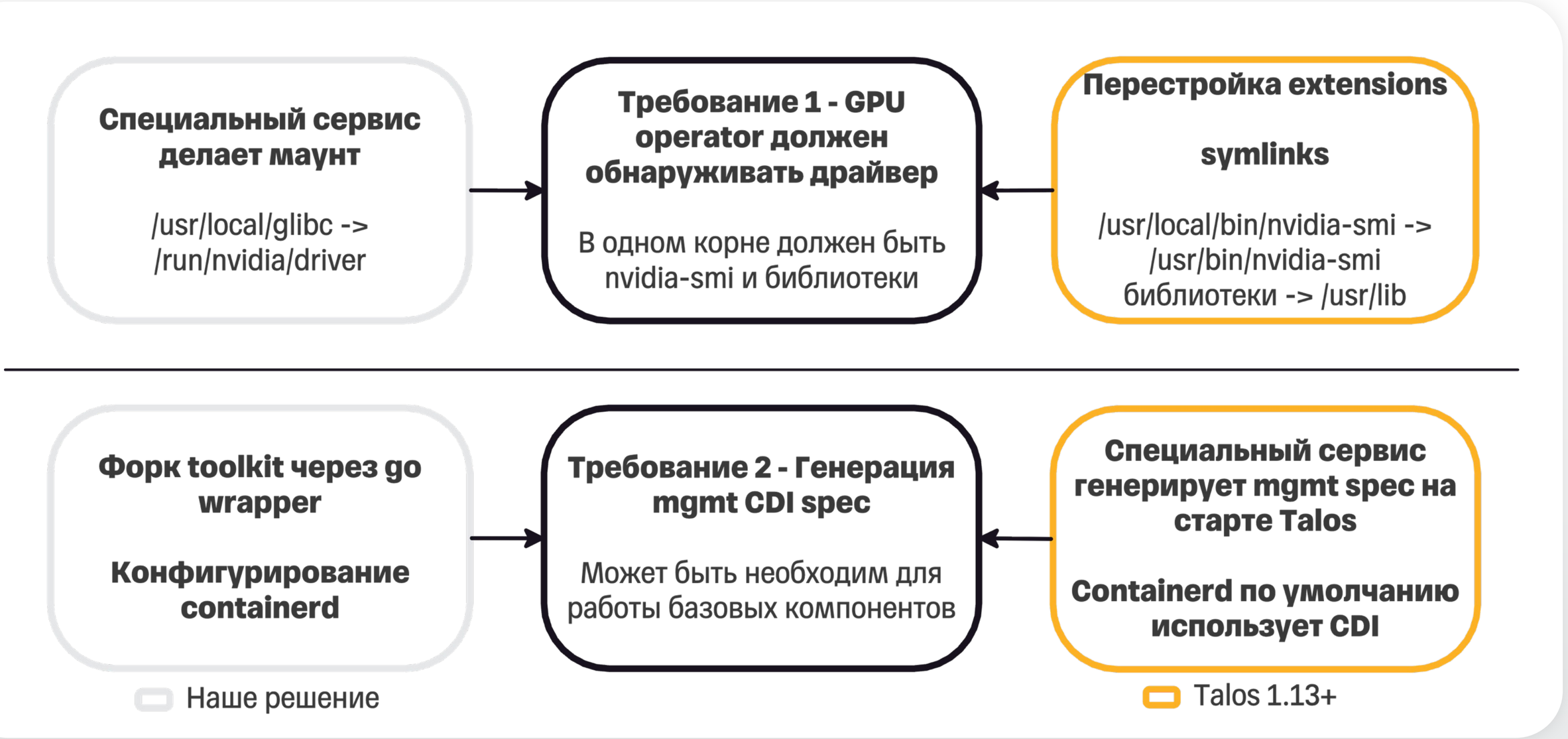
Talos 1.13+


Специальный сервис делает маунт
/usr/local/glibc -> /run/nvidia/driver
Требование 1 - GPU operator должен обнаруживать драйвер
В одном корне должен быть nvidia-smi и библиотеки
Перестройка extensions
symlinks
/usr/local/bin/nvidia-smi -> /usr/bin/nvidia-smi
библиотеки -> /usr/lib
Форк toolkit через go wrapper
Требование 2 - Генерация mgmt CDI spec
Может быть необходим для работы базовых компонентов
Специальный сервис генерирует mgmt spec на старте Talos
Наше решение
Talos 1.13+

Talos 1.13+


Специальный сервис делает маунт
/usr/local/glibc -> /run/nvidia/driver
Требование 1 - GPU operator должен обнаруживать драйвер
В одном корне должен быть nvidia-smi и библиотеки
Перестройка extensions
symlinks
/usr/local/bin/nvidia-smi -> /usr/bin/nvidia-smi
библиотеки -> /usr/lib
Форк toolkit через go wrapper
Конфигурирование containerd
Требование 2 - Генерация mgmt CDI spec
Может быть необходим для работы базовых компонентов
Специальный сервис генерирует mgmt spec на старте Talos
Наше решение
Talos 1.13+

Talos 1.13+


Специальный сервис делает маунт
/usr/local/glibc -> /run/nvidia/driver
Требование 1 - GPU operator должен обнаруживать драйвер
В одном корне должен быть nvidia-smi и библиотеки
Перестройка extensions
symlinks
/usr/local/bin/nvidia-smi -> /usr/bin/nvidia-smi
библиотеки -> /usr/lib
Форк toolkit через go wrapper
Конфигурирование containerd
Требование 2 - Генерация mgmt CDI spec
Может быть необходим для работы базовых компонентов
Специальный сервис генерирует mgmt spec на старте Talos
Containerd по умолчанию использует CDI
Наше решение
Talos 1.13+

Talos 1.13+ Выводы

Плюсы

- Используются upstream версии компонентов

Минусы

- Для обновления Container Toolkit необходимо обновлять Talos
- GPU Operator ничего не знает о сервисе генерации CDI



Заглянем в будущее

Хочется вот так

Плюсы

- Используются upstream версии компонентов
- Для обновления Container Toolkit НЕ нужно обновлять Talos

Минусы

Node Resource Interface

Node Resource Interface



Для mgmt подов NVIDIA

Дает служебным подам доступ к GPU без выделения GPU через Kubernetes

Node Resource Interface

Для mgmt подов NVIDIA

Дает служебным подам доступ к GPU без выделения GPU через Kubernetes

Связка K8S и CDI

NRI превращает аннотации pod в запрос на подключение CDI-устройств

Node Resource Interface

Для mgmt подов NVIDIA

Дает служебным подам доступ к GPU без выделения GPU через Kubernetes

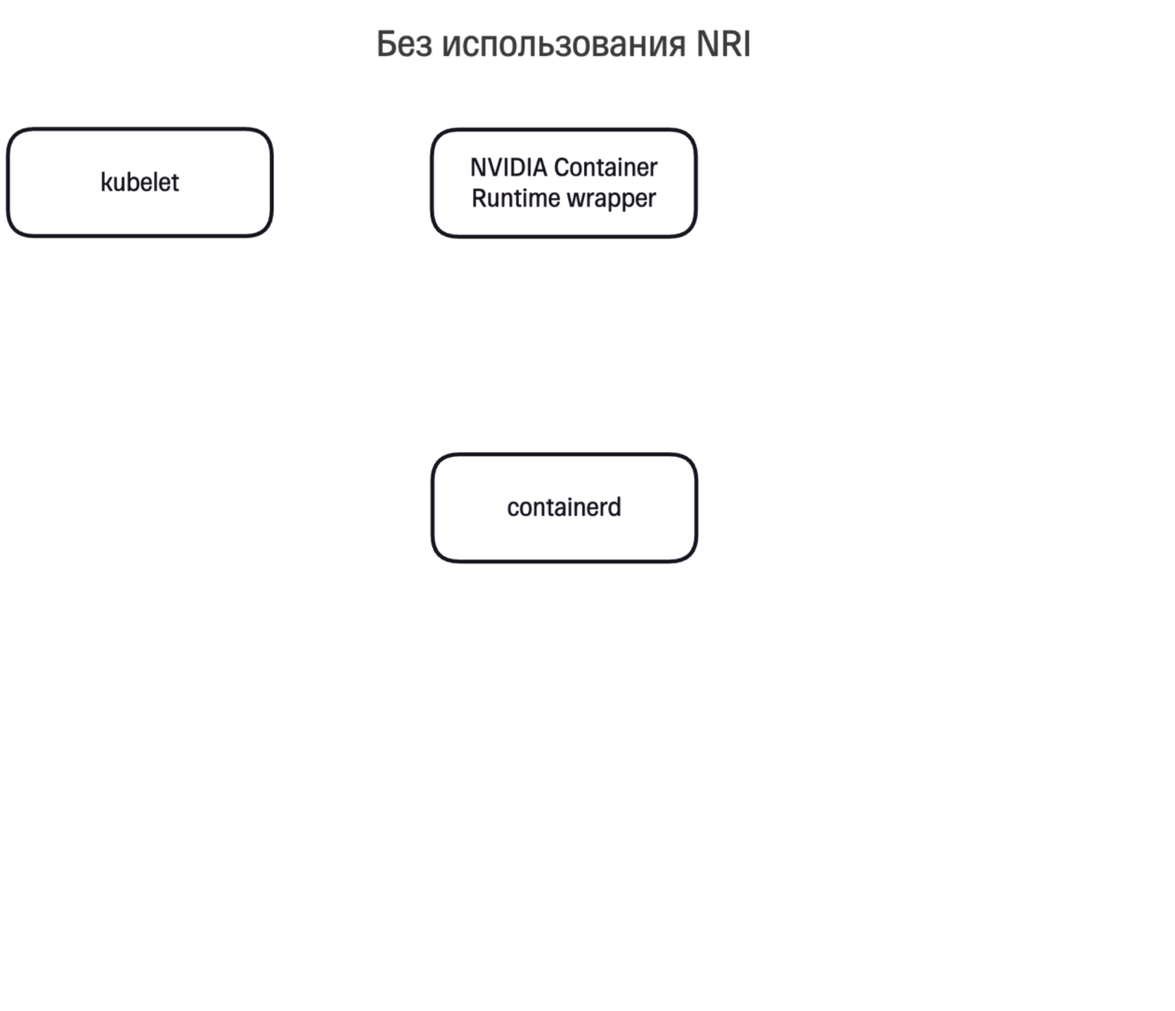
Связка K8S и CDI

NRI превращает аннотации pod в запрос на подключение CDI-устройств

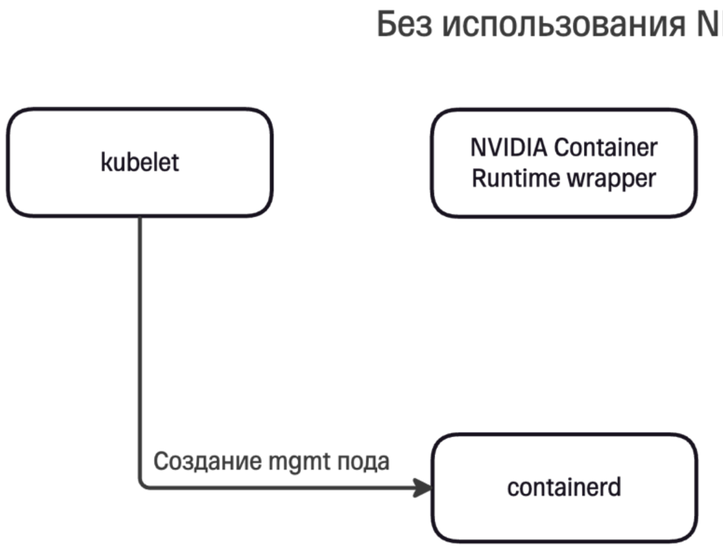
Без nvidia RuntimeClass

GPU подключаются через NRI + CDI, без отдельного NVIDIA runtime

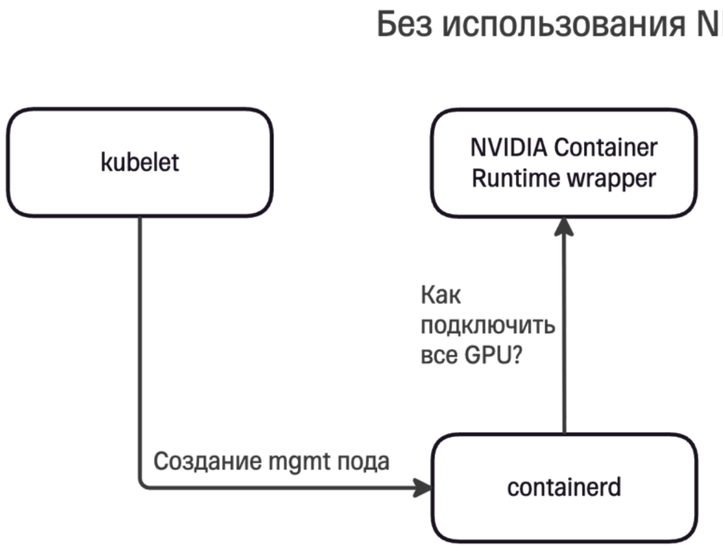
NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
containerd

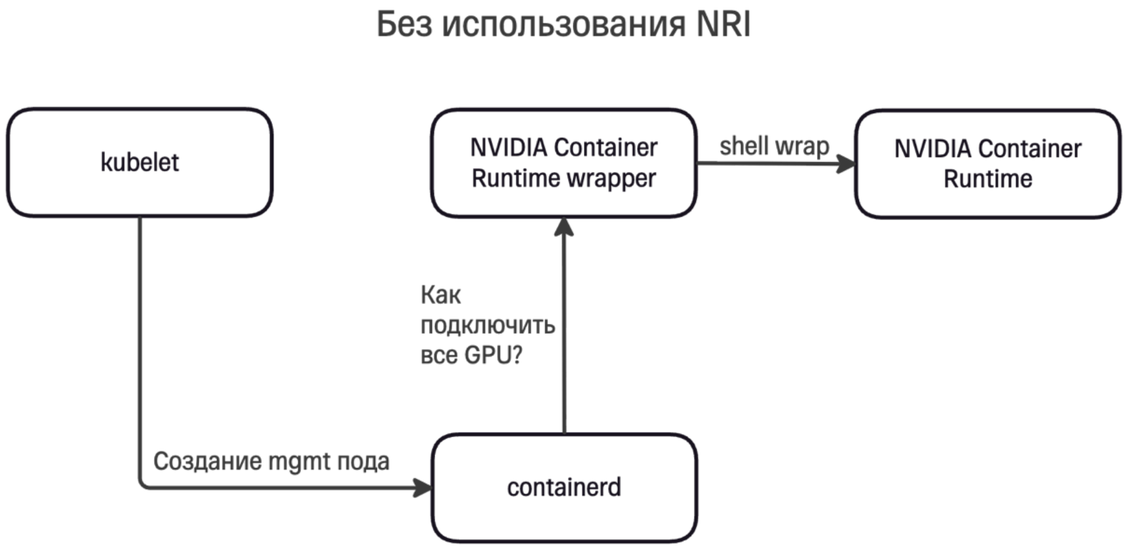
NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
Создание mgmt пода
containerd

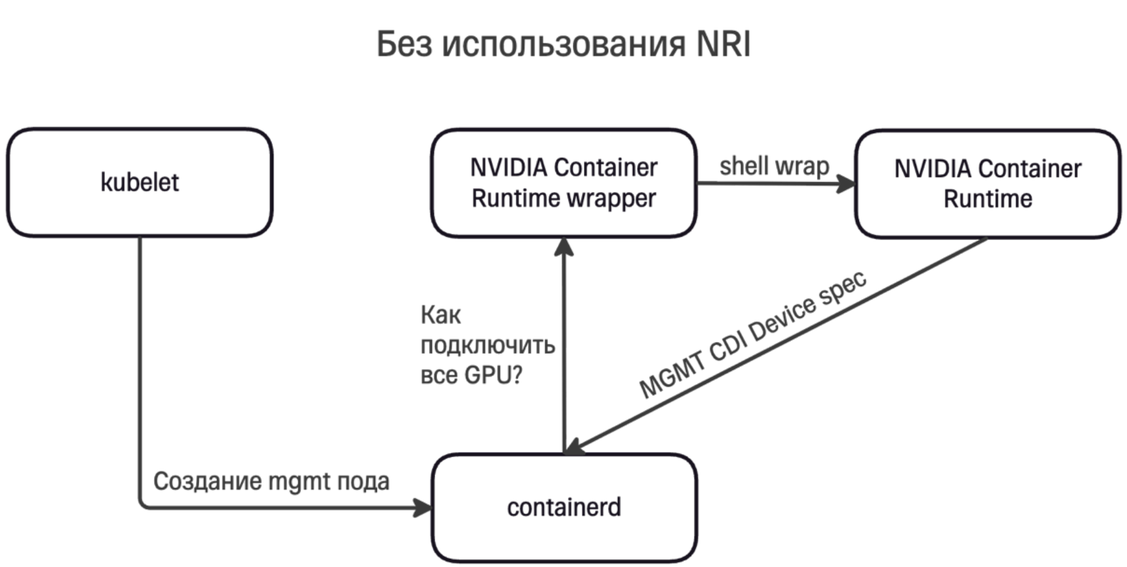
NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
Как подключить все GPU?
Создание mgmt пода
containerd

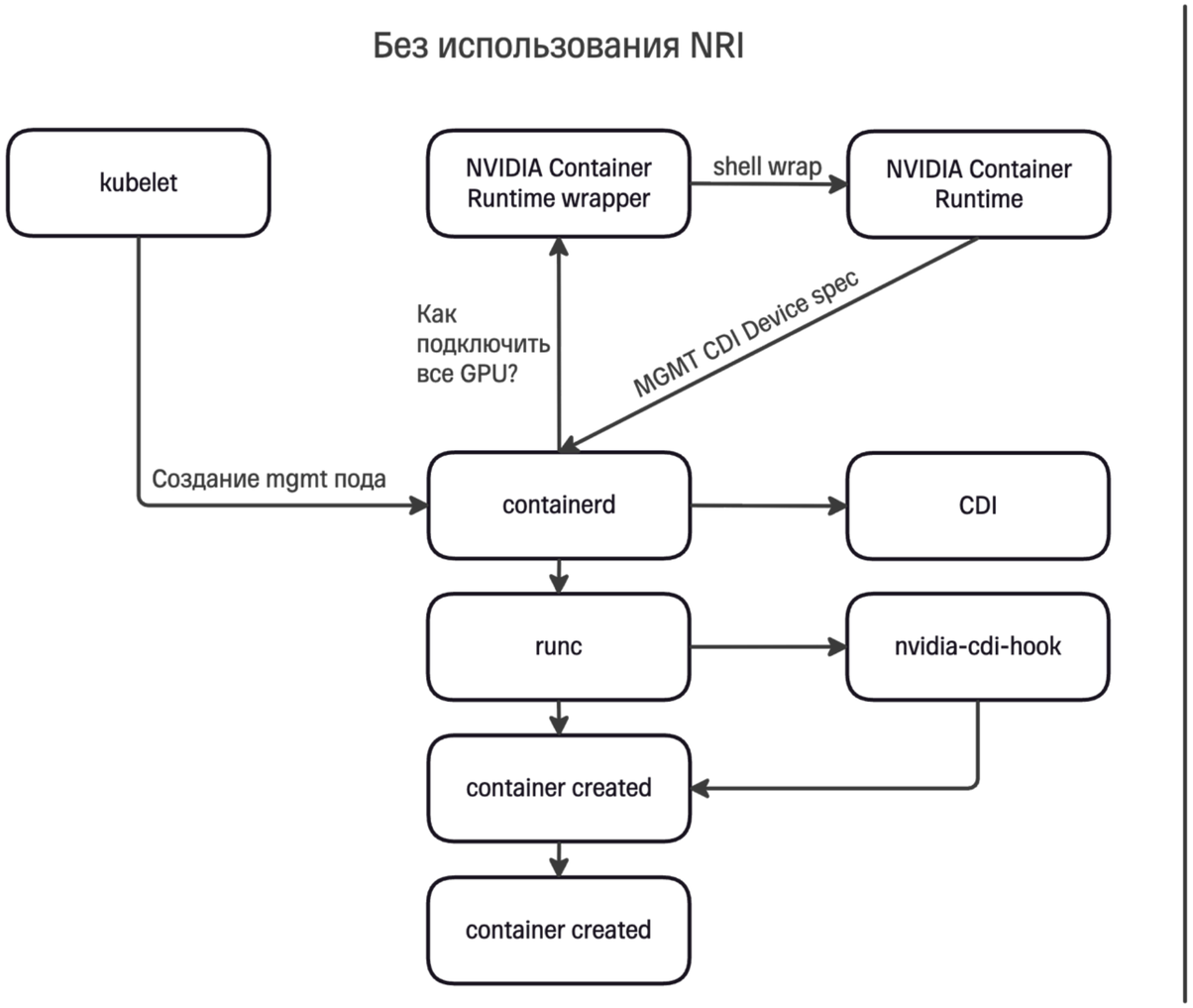
NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
Создание mgmt пода
containerd

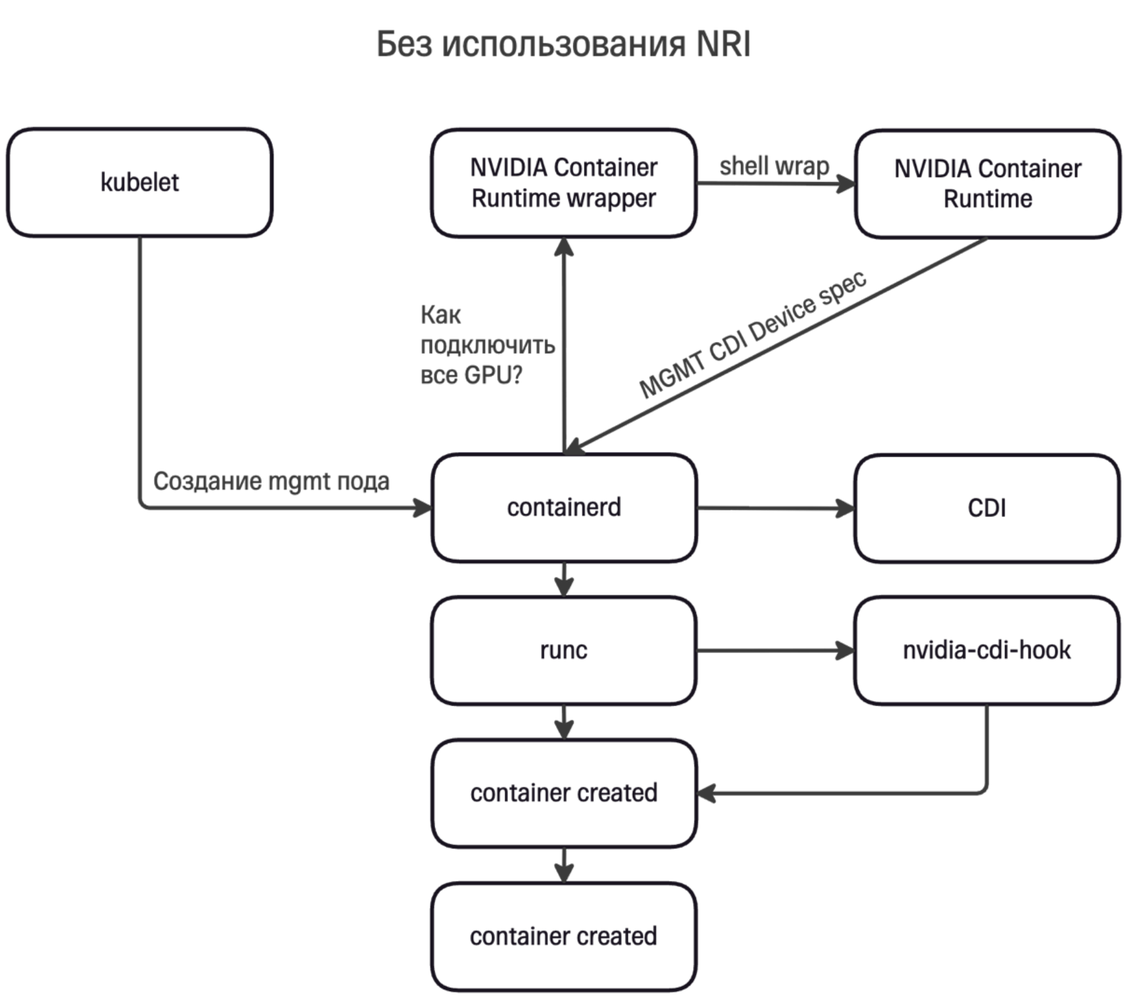
NRI и GPU

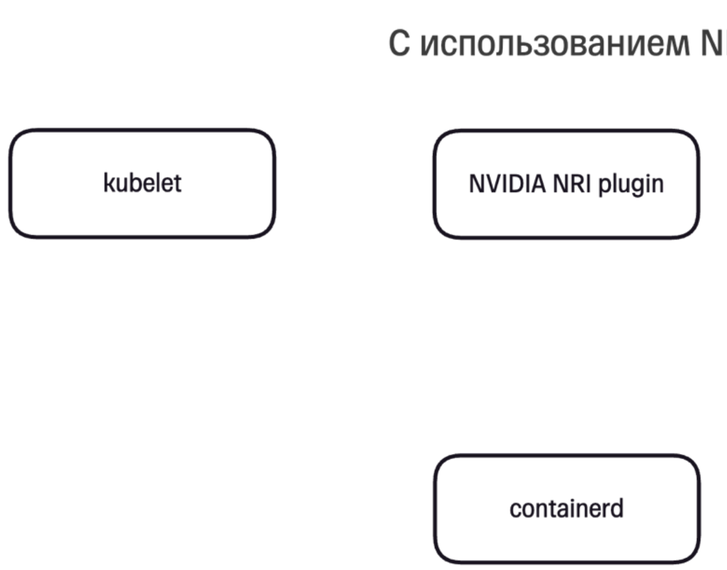

Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
MGMT CDI Device spec
Создание mgmt пода
containerd

NRI и GPU

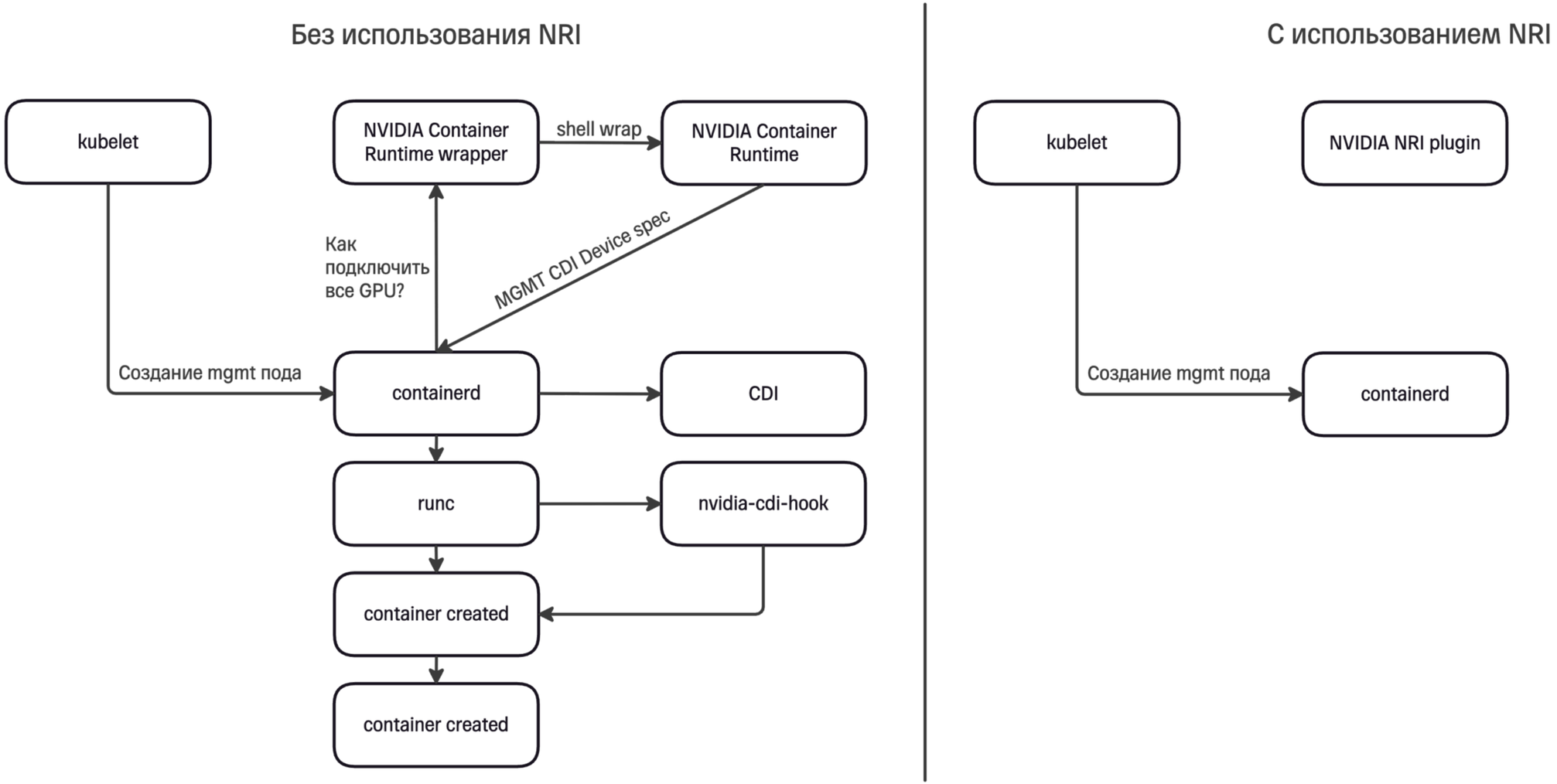

Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
MGMT CDI Device spec
Создание mgmt пода
containerd
CDI
runc
nvidia-cdi-hook
container created
container created

NRI и GPU

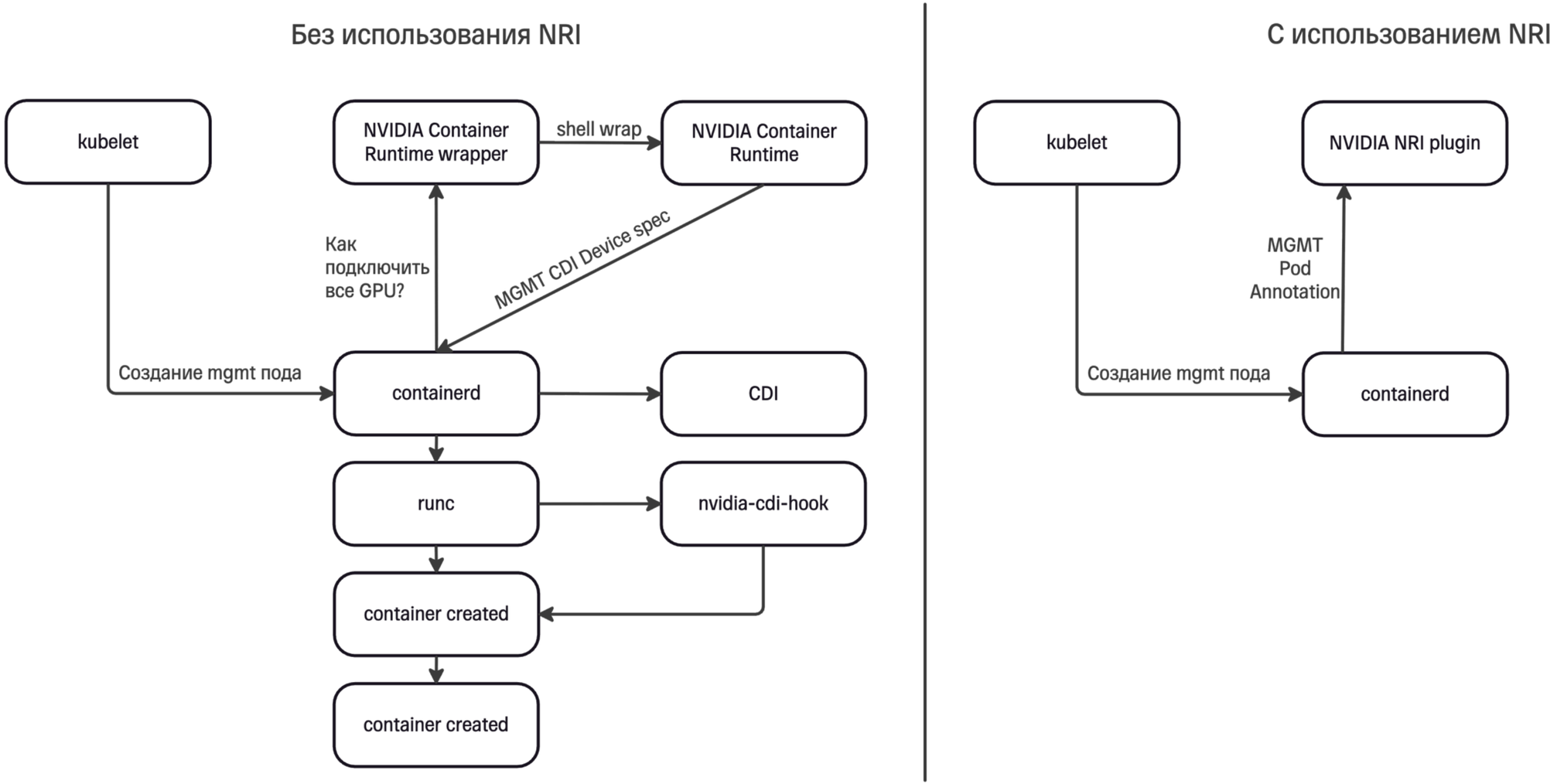

Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
MGMT CDI Device spec
Создание mgmt пода
containerd
CDI
runc
nvidia-cdi-hook
container created
container created


С использованием NRI
kubelet
NVIDIA NRI plugin
containerd

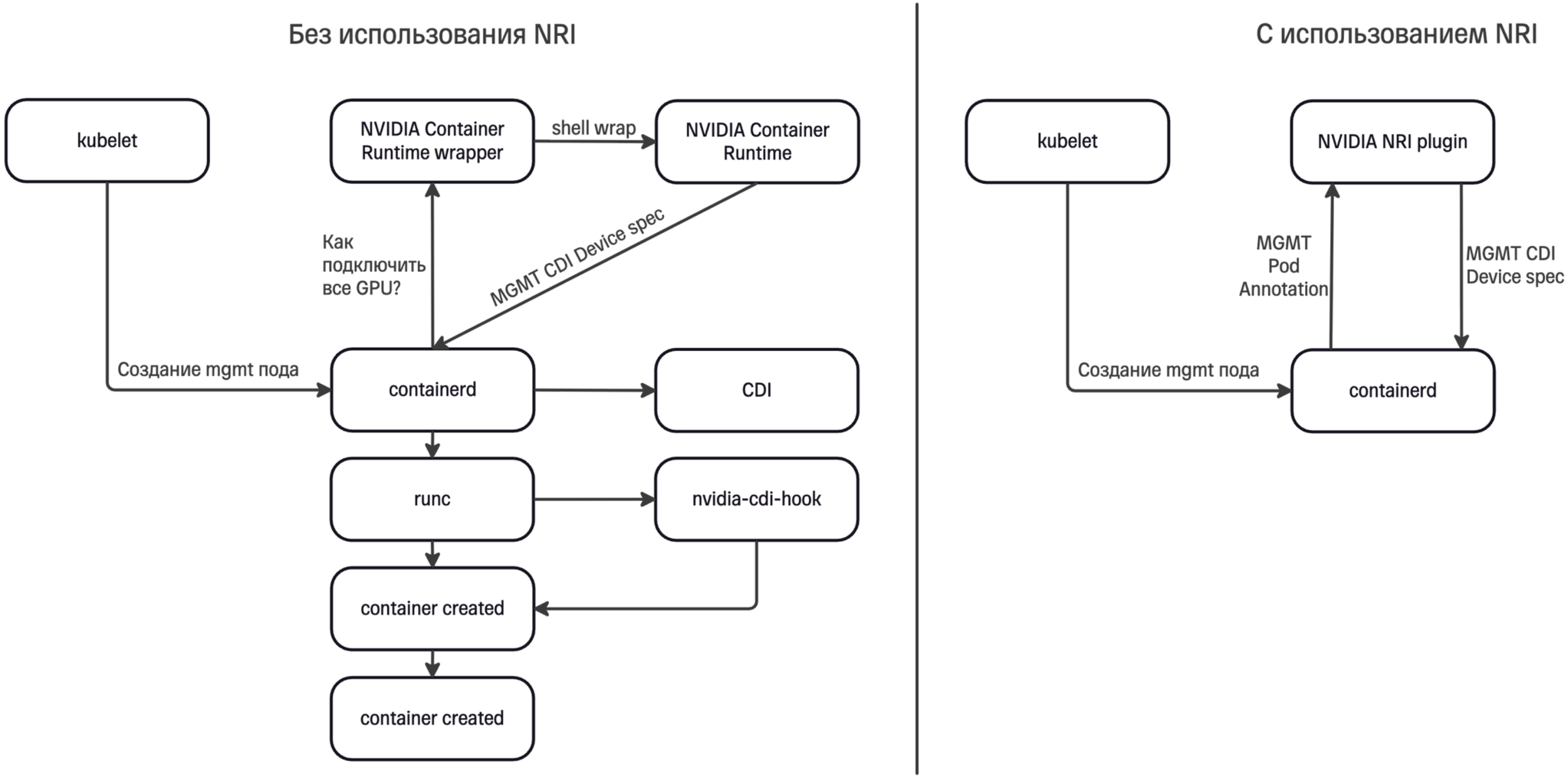
NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
MGMT CDI Device spec
Создание mgmt пода
containerd
CDI
runc
nvidia-cdi-hook
container created
container created
С использованием NRI
kubelet
NVIDIA NRI plugin
Создание mgmt пода
containerd

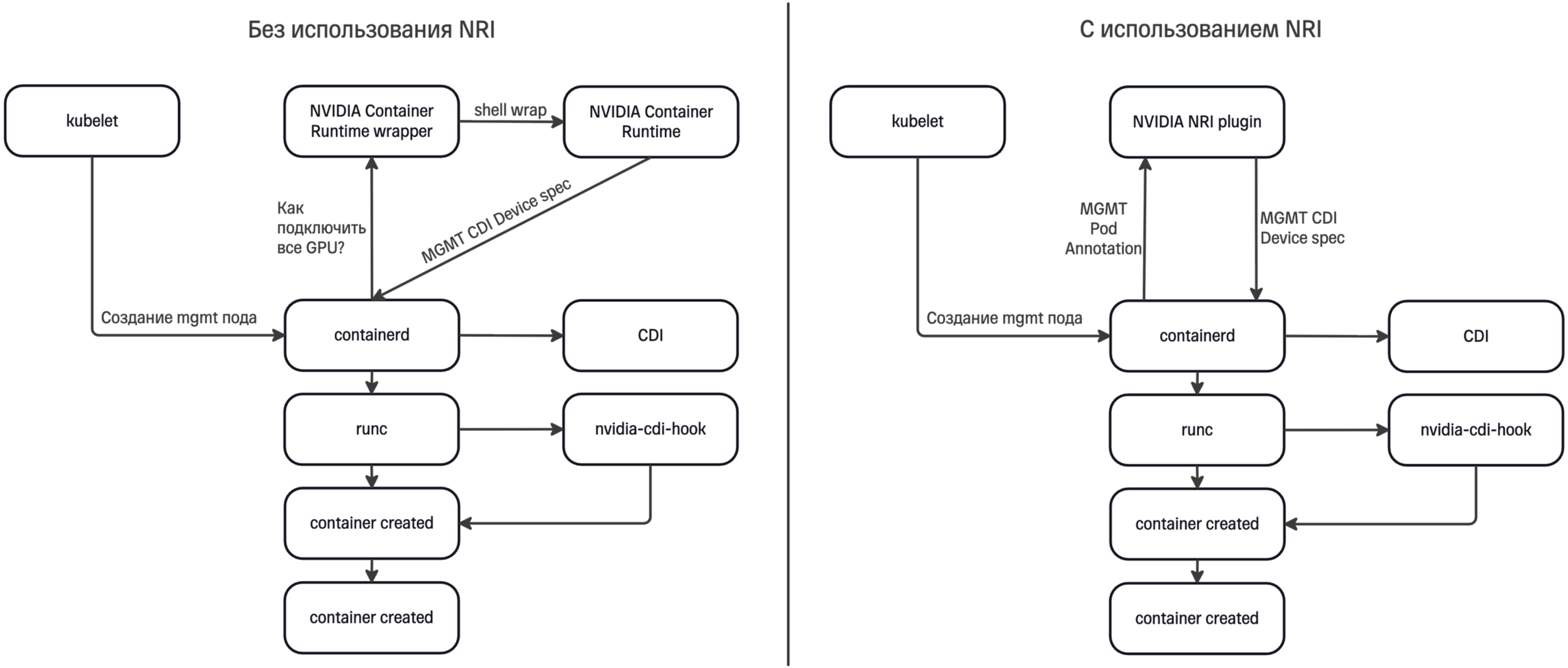
NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
MGMT CDI Device spec
Создание mgmt пода
containerd
CDI
runc
nvidia-cdi-hook
container created
container created
С использованием NRI
kubelet
NVIDIA NRI plugin
MGMT Pod Annotation
Создание mgmt пода
containerd

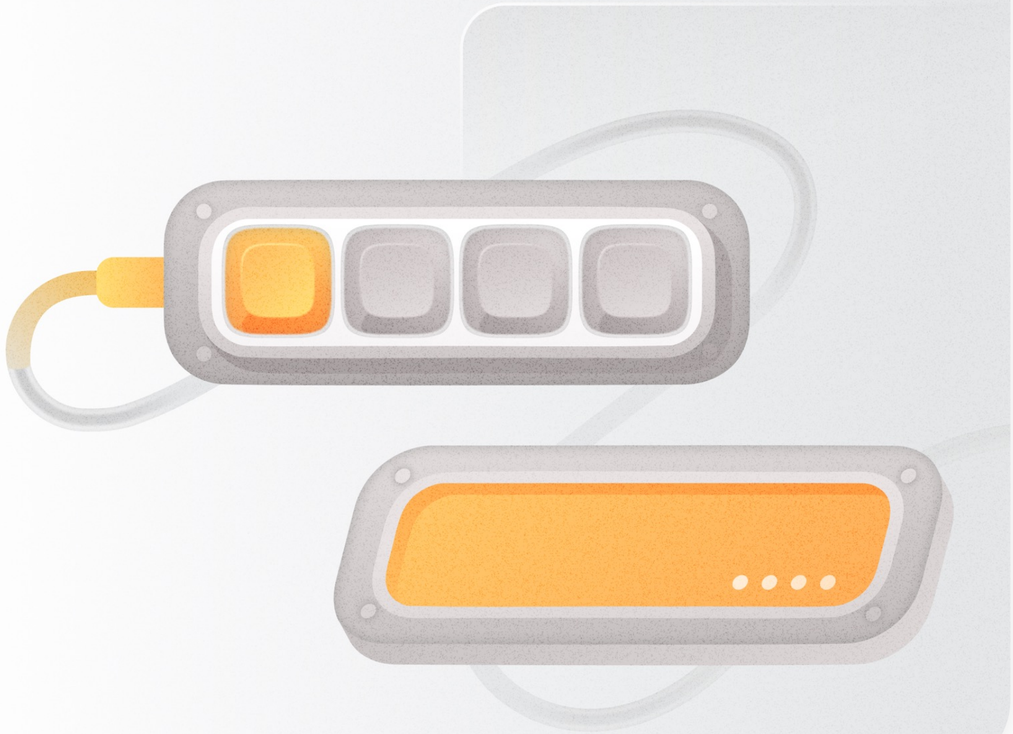
NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
MGMT CDI Device spec
Создание mgmt пода
containerd
CDI
runc
nvidia-cdi-hook
container created
container created
С использованием NRI
kubelet
NVIDIA NRI plugin
MGMT Pod Annotation
MGMT CDI Device spec
Создание mgmt пода
containerd

NRI и GPU


Без использования NRI
kubelet
NVIDIA Container Runtime wrapper
shell wrap
NVIDIA Container Runtime
Как подключить все GPU?
MGMT CDI Device spec
Создание mgmt пода
containerd
CDI
runc
nvidia-cdi-hook
container created
container created
С использованием NRI
kubelet
NVIDIA NRI plugin
MGMT Pod Annotation
MGMT CDI Device spec
Создание mgmt пода
containerd
CDI
runc
nvidia-cdi-hook
container created
container created

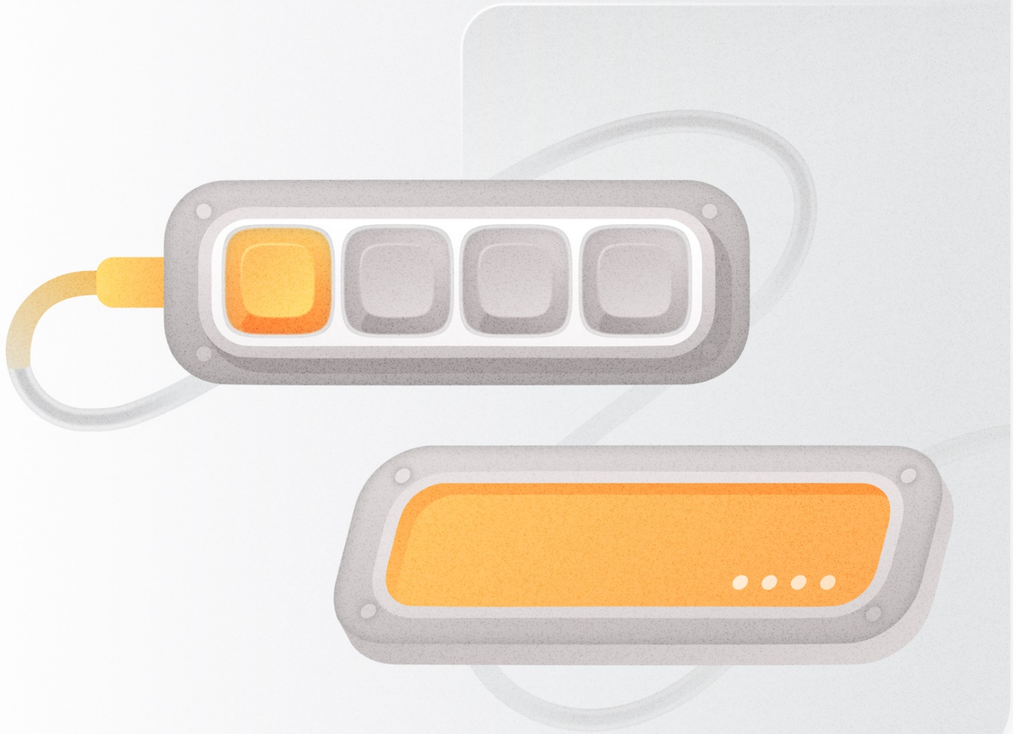
NRI и Talos



NRI by default

- В Talos 1.14+ NRI включен по умолчанию
- В Talos < 1.13 нужно включать самостоятельно

NRI и Talos



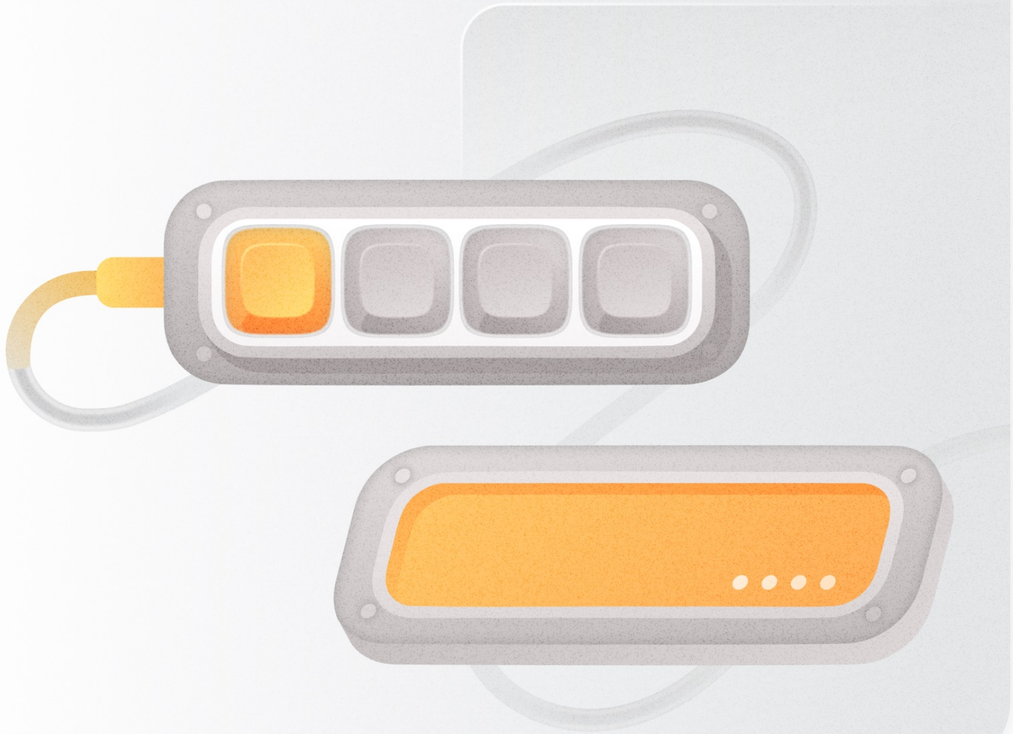
NRI by default

- В Talos 1.14+ NRI включен по умолчанию
- В Talos < 1.13 нужно включать самостоятельно

NRI и host Toolkit

NRI плагин официально поддерживается только в составе Container Toolkit DaemonSet

NRI и Talos



NRI by default

- В Talos 1.14+ NRI включен по умолчанию
- В Talos < 1.13 нужно включать самостоятельно

NRI и host Toolkit

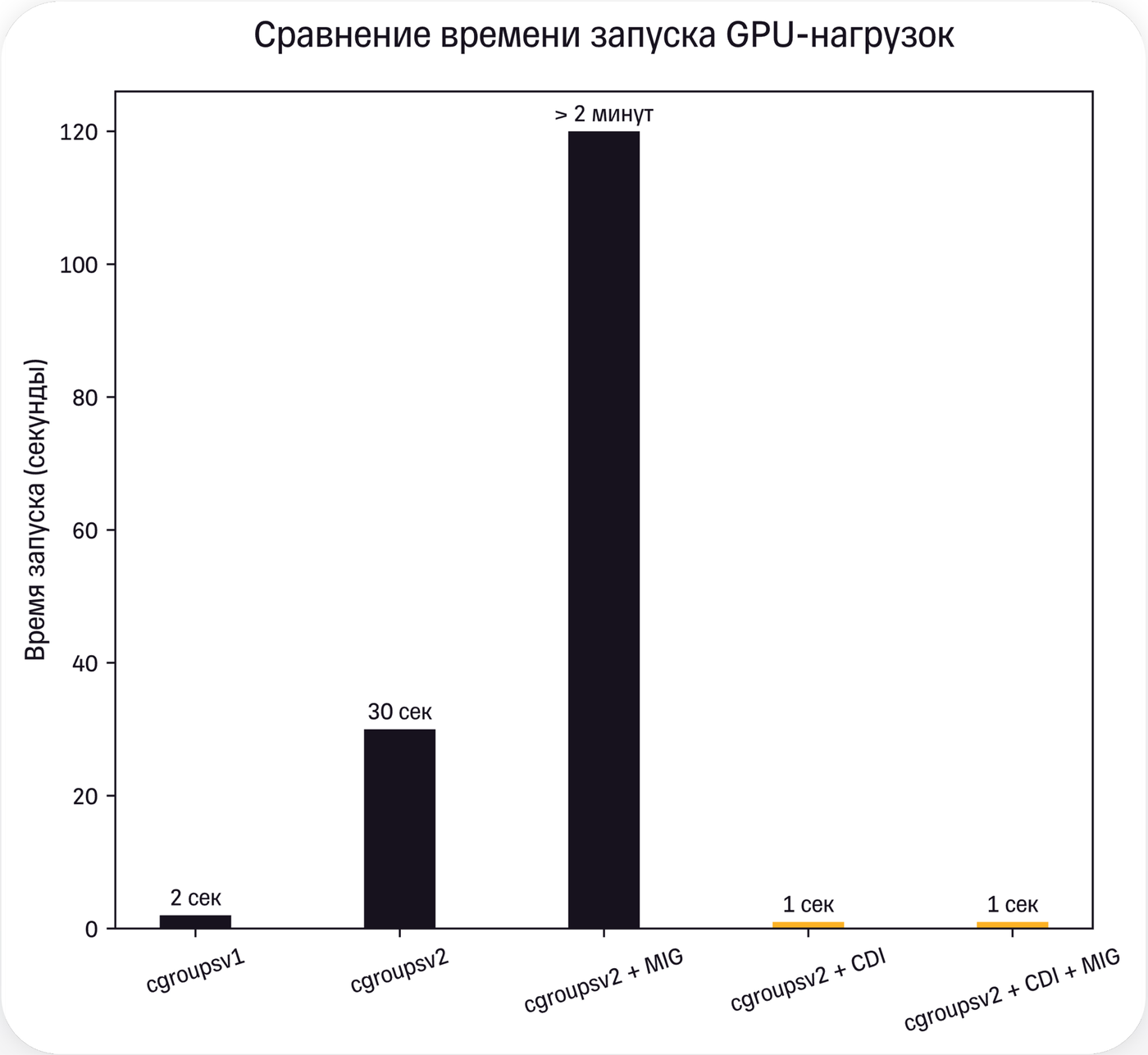
NRI плагин официально поддерживается только в составе Container Toolkit DaemonSet

Container Toolkit Extension

- В Talos 1.15 из Container Toolkit extension будет убран Container Toolkit
- Container Toolkit будет устанавливаться как DaemonSet

Промежуточные итоги

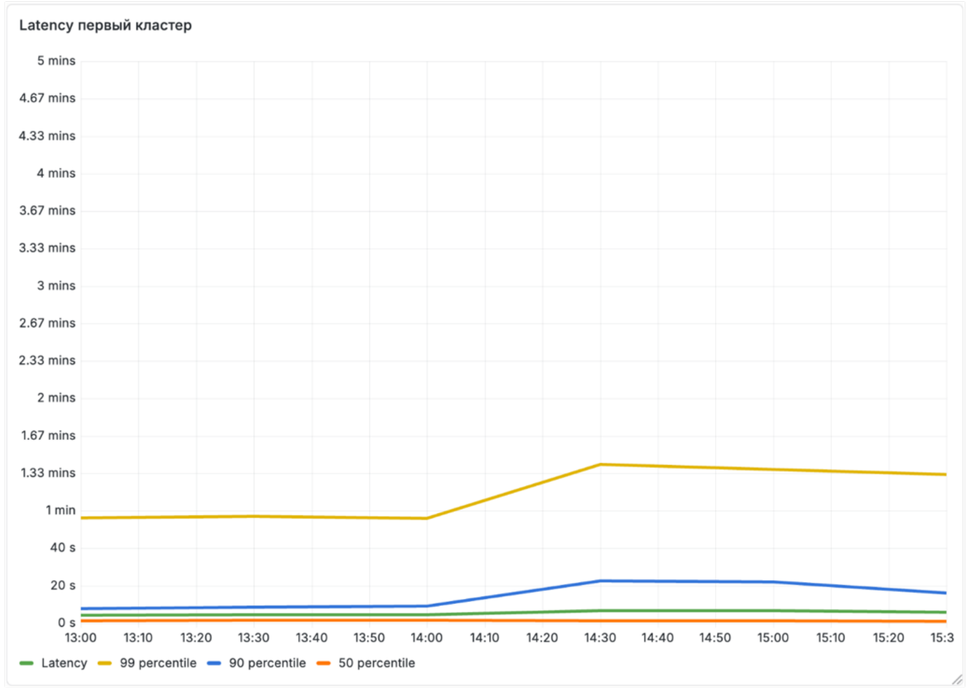
- Разблокировали использование MIG
- Запустили NVIDIA GPU Operator
- Запустились на cgroupsv2



Проблема 2 – Высокий latency

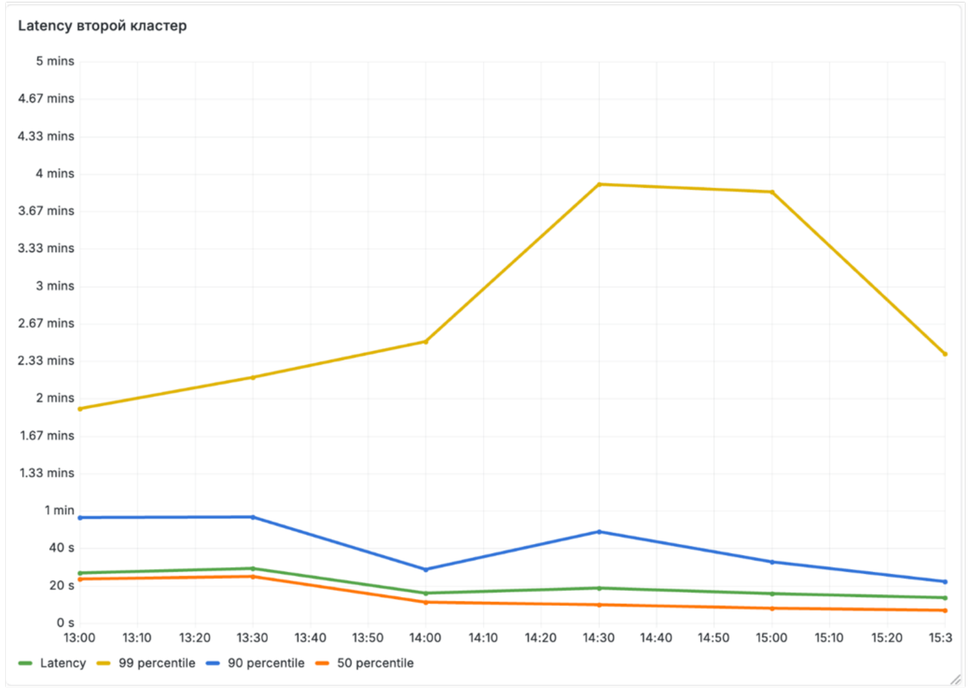
Время генерации ответа LLM моделью

Первый кластер



Production трафик

Второй кластер

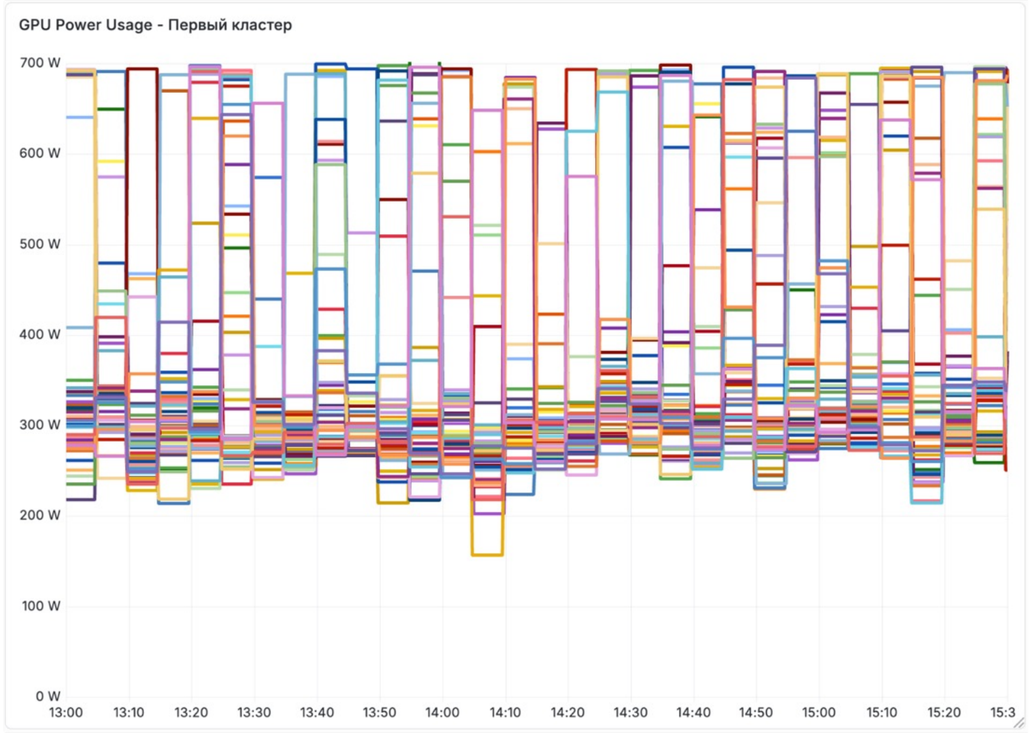


Зеркалирование 1/3 трафика

Показатели

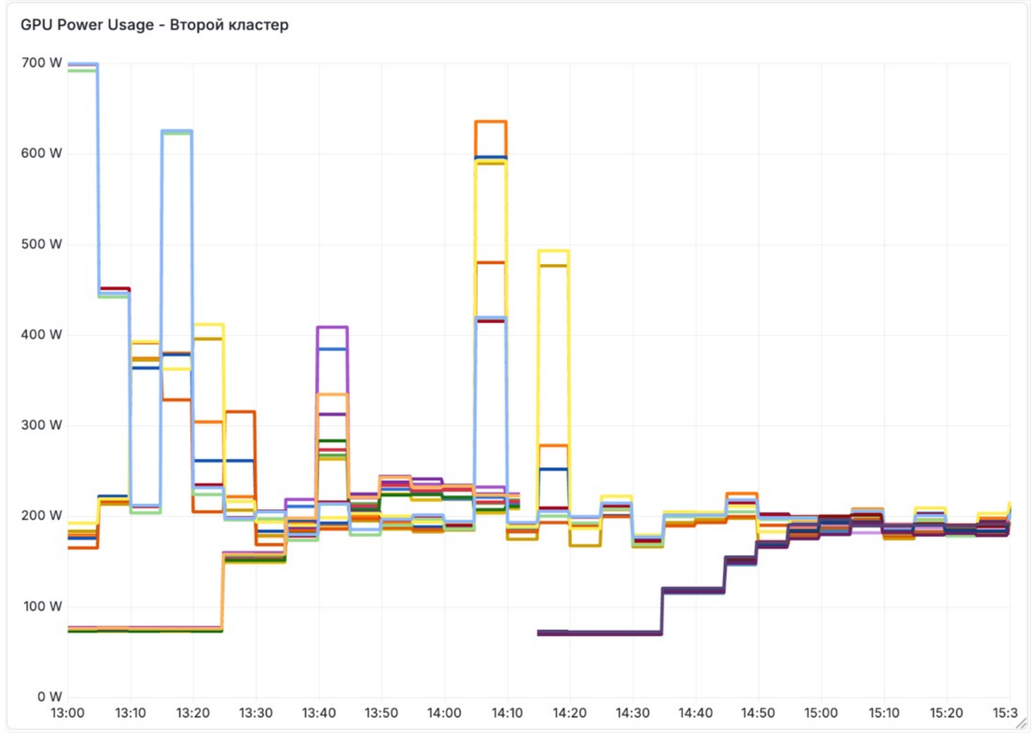
Потребление питания GPU в кластере

Первый кластер



300w с частыми пиками до 700w

Второй кластер

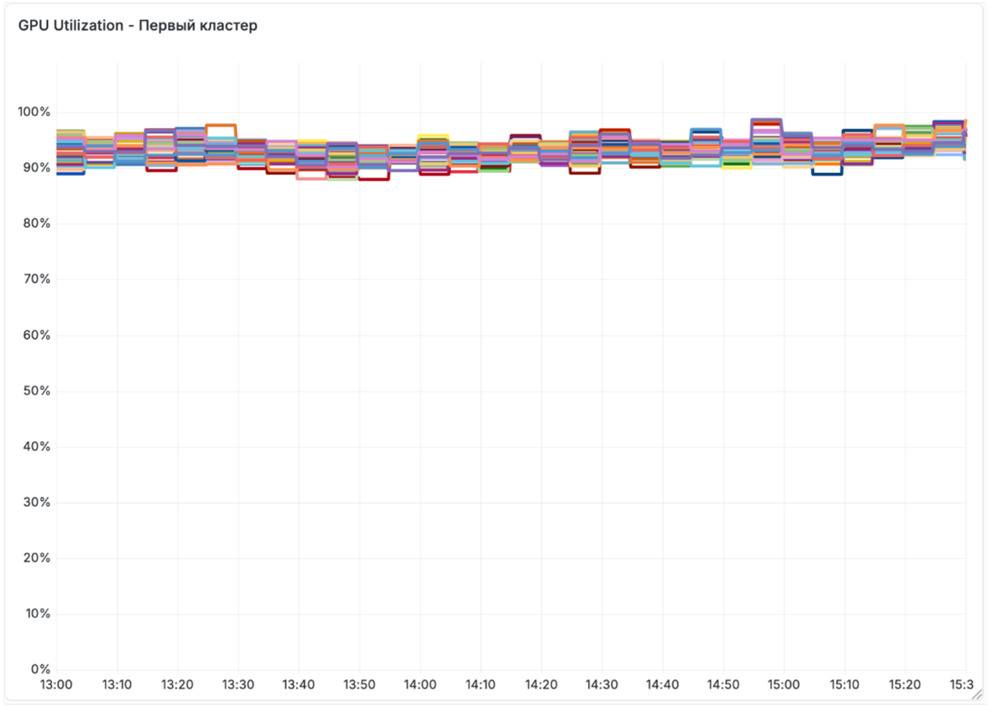


180w с редкими пиками до 700w

Показатели

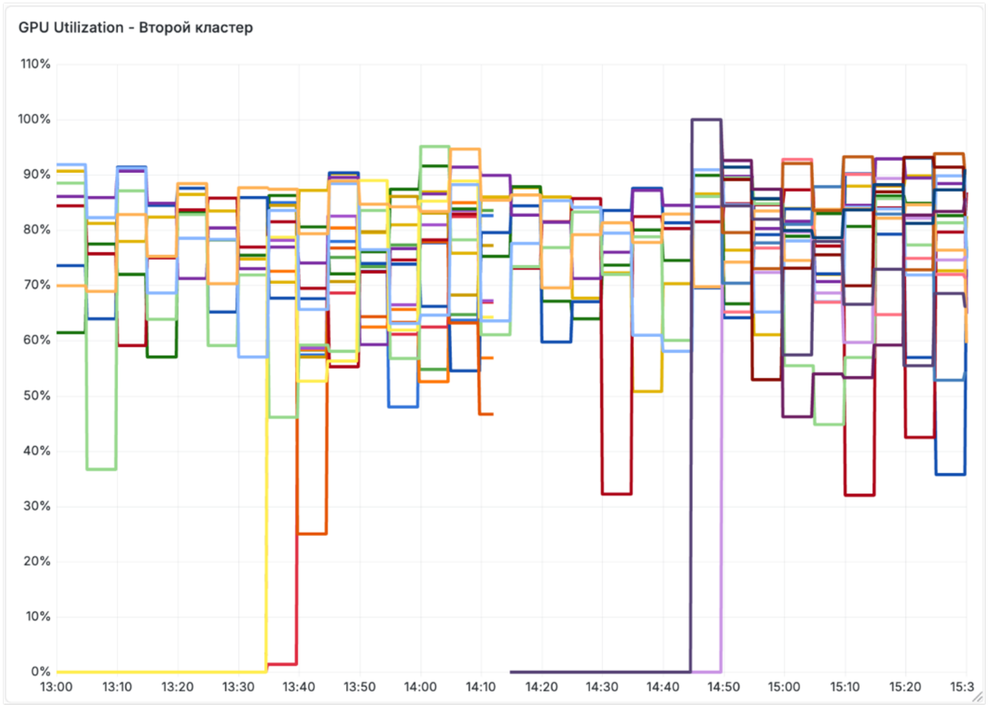
Утилизация GPU в кластере

Первый кластер



Утилизация в районе 92%

Второй кластер



Нестабильная утилизация

Поиск источника

GPU

01

Поиск источника

GPU

GPU Burner показывает отличные результаты

01

Поиск источника

GPU

GPU Burner показывает отличные результаты

01

Проблемные связки GPU

02

Поиск источника

GPU

GPU Burner показывает отличные результаты

01

Проблемные связки GPU

- Все NVLink видны программно
- Все пары видеокарт показывают одинаковую производительность

02

Поиск источника

GPU

GPU Burner показывает отличные результаты

01

Проблемные связки GPU

- Все NVLink видны программно
- Все пары видеокарт показывают одинаковую производительность

02

BIOS

03

Поиск источника

GPU

GPU Burner показывает отличные результаты

01

Проблемные связки GPU

- Все NVLink видны программно
- Все пары видеокарт показывают одинаковую производительность

02

BIOS

Проверили:

- Версии прошивок BIOS, CPLD, PSU
- Конфигурацию BIOS

03

Поиск источника

GPU

GPU Burner показывает отличные результаты

01

Проблемные связки GPU

- Все NVLink видны программно
- Все пары видеокарт показывают одинаковую производительность

02

BIOS

Проверили:

- Версии прошивок BIOS, CPLD, PSU
- Конфигурацию BIOS

03

Физические особенности

04

Поиск источника

GPU

GPU Burner показывает отличные результаты

01

Проблемные связки GPU

- Все NVLink видны программно
- Все пары видеокарт показывают одинаковую производительность

02

BIOS

Проверили:

- Версии прошивок BIOS, CPLD, PSU
- Конфигурацию BIOS

03

Физические особенности

- Переусадили/почистили все GPU
- Проверили все NVLink

04

Разделяемся

В ходе попыток исправить проблему получили один исправный сервер

Бенчмарк multi-GPU нагрузки

- Мы не можем нормально тестировать гипотезы без пользовательской нагрузки
- Нет референсных значений

Разделяемся

В ходе попыток исправить проблему получили один исправный сервер

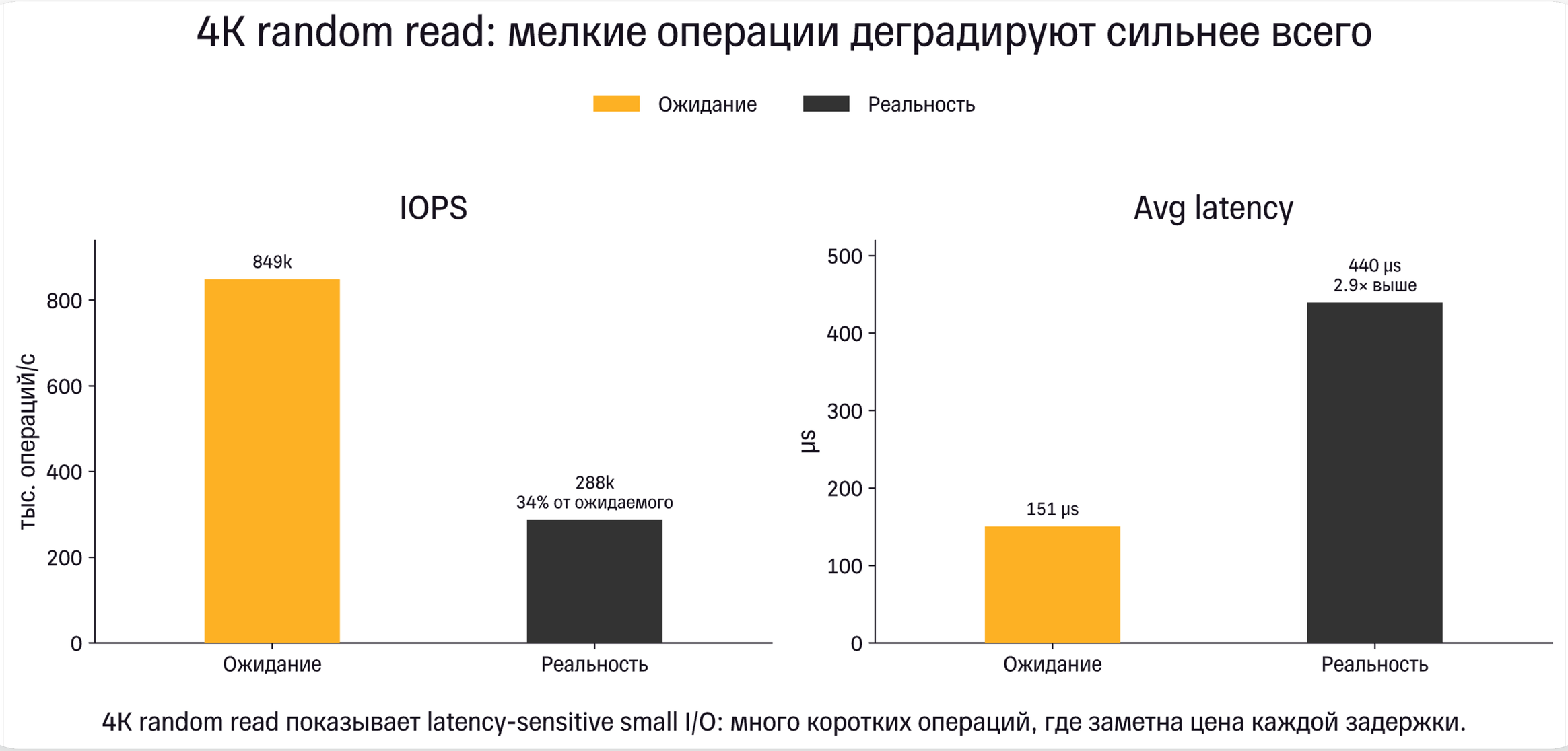
Бенчмарк multi-GPU нагрузки

- Мы не можем нормально тестировать гипотезы без пользовательской нагрузки
- Нет референсных значений

Проверка дисков

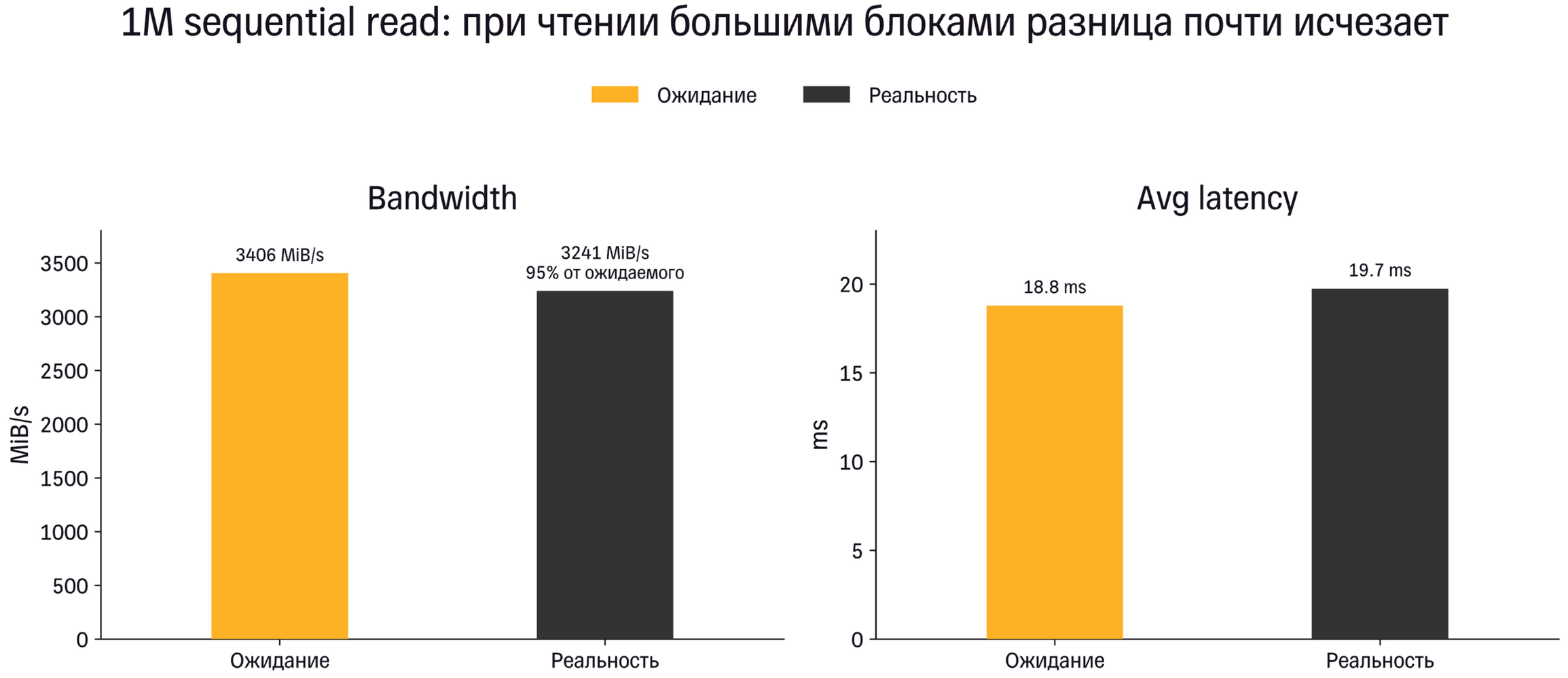
- Большие образы пулятся в 3 раза дольше чем на stage кластере
- Диски на prod кластере значительно быстрее

Проверка дисков – 4k random read



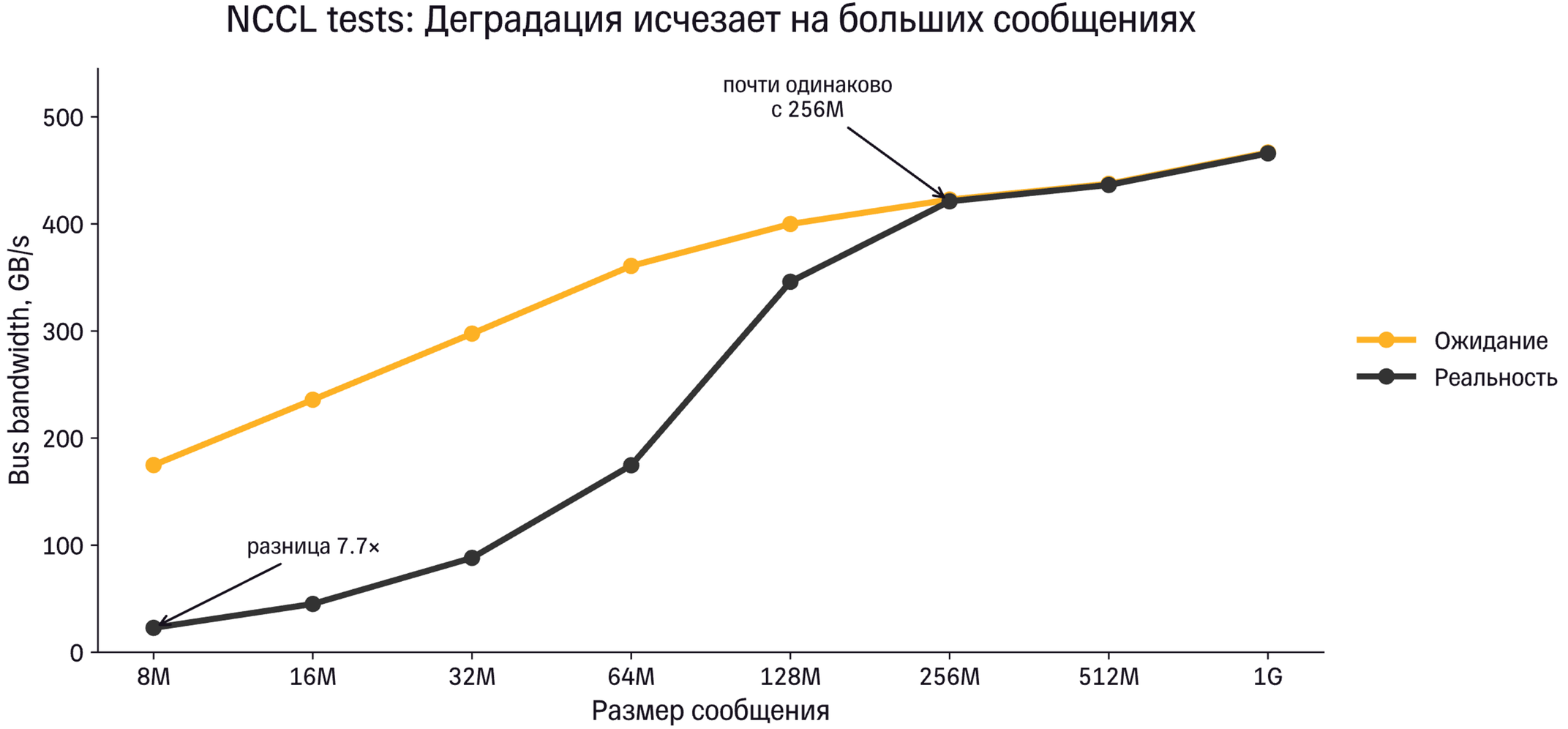
4K random read показывает latency-sensitive small I/O: много коротких операций, где заметна цена каждой задержки.

Проверка дисков – 1M sequential read



1M sequential read показывает peak throughput: при последовательном чтении большими блоками оба сервера выходят почти на одинаковый bandwidth

Бенчмарк multi-GPU нагрузки



Малые collective-операции деградируют значительно сильнее крупных передач данных.

Что видим?

Что видим?



PCIe

NVME SSD и GPU

подключаются по PCIe

Что видим?



PCIe

NVME SSD и GPU подключаются по PCIe



Только small batch

Проблема видна только на малых данных. Эти данные быстро обрабатываются SSD и GPU

Что видим?

PCIe

NVME SSD и GPU подключаются по PCIe

Только small batch

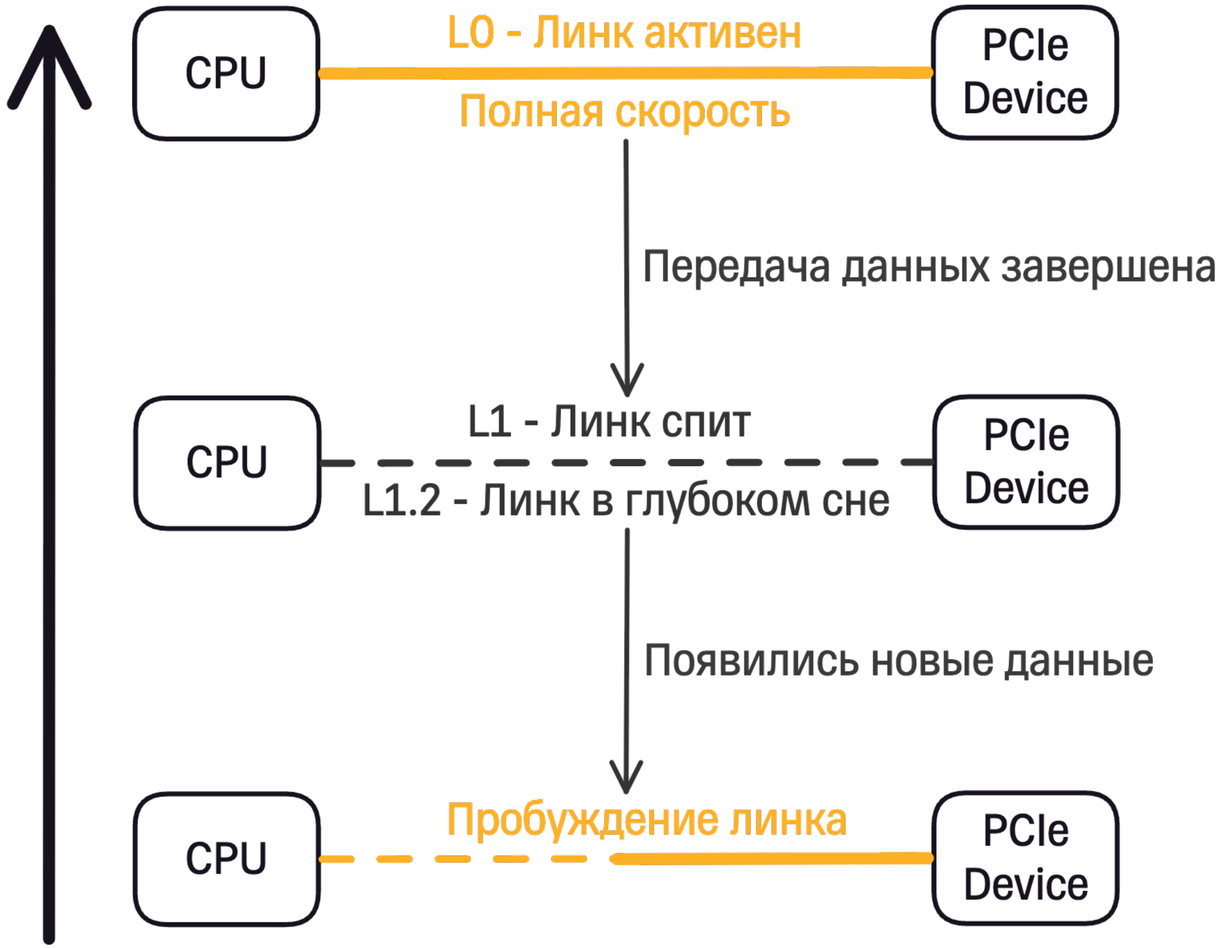
Проблема видна только на малых данных. Эти данные быстро обрабатываются SSD и GPU

Бенчмарки идеальны

При 100% потреблении устройств проблем нет

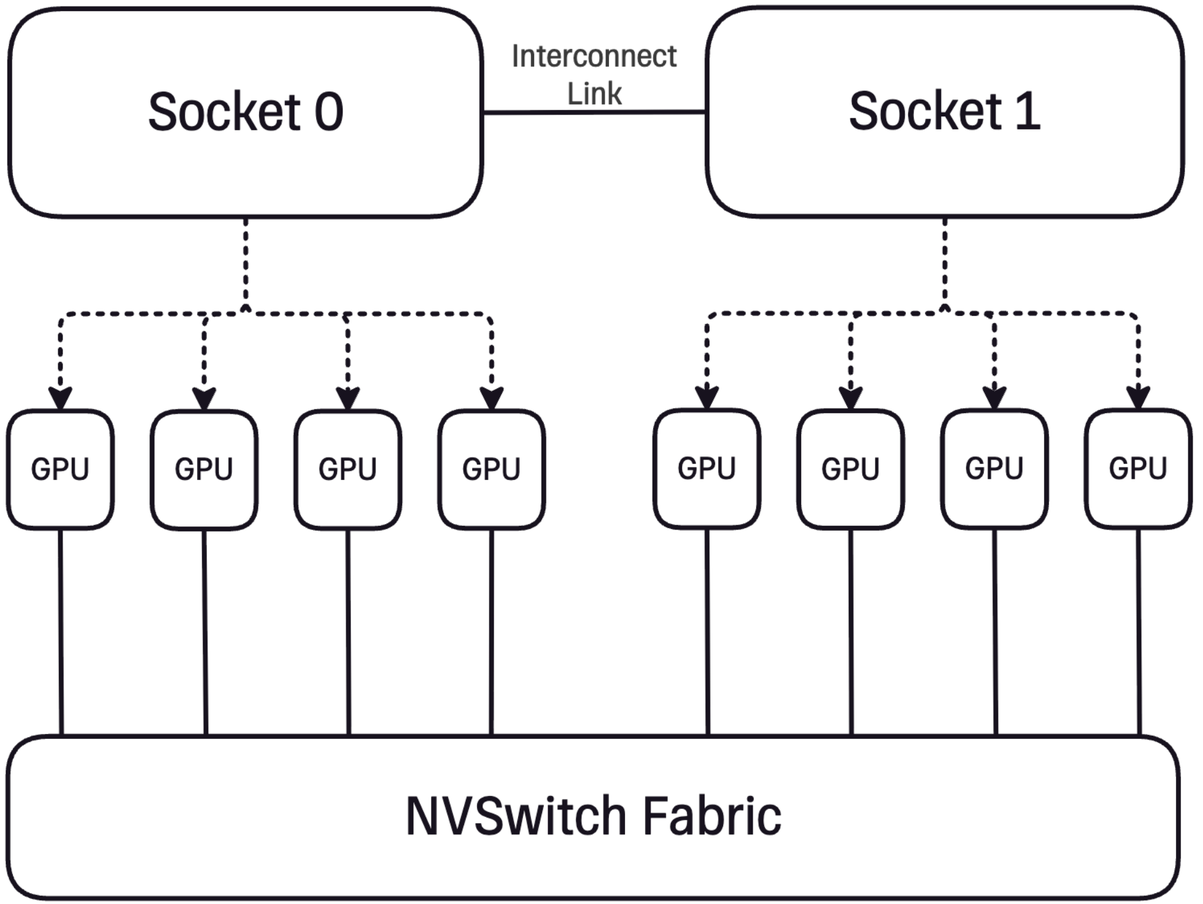
PCIe ASPM

- Малые данные быстро обрабатываются PCIe устройствами
- Power Manager успевает погрузить линки в сон
- Каждое такое пробуждение – задержки



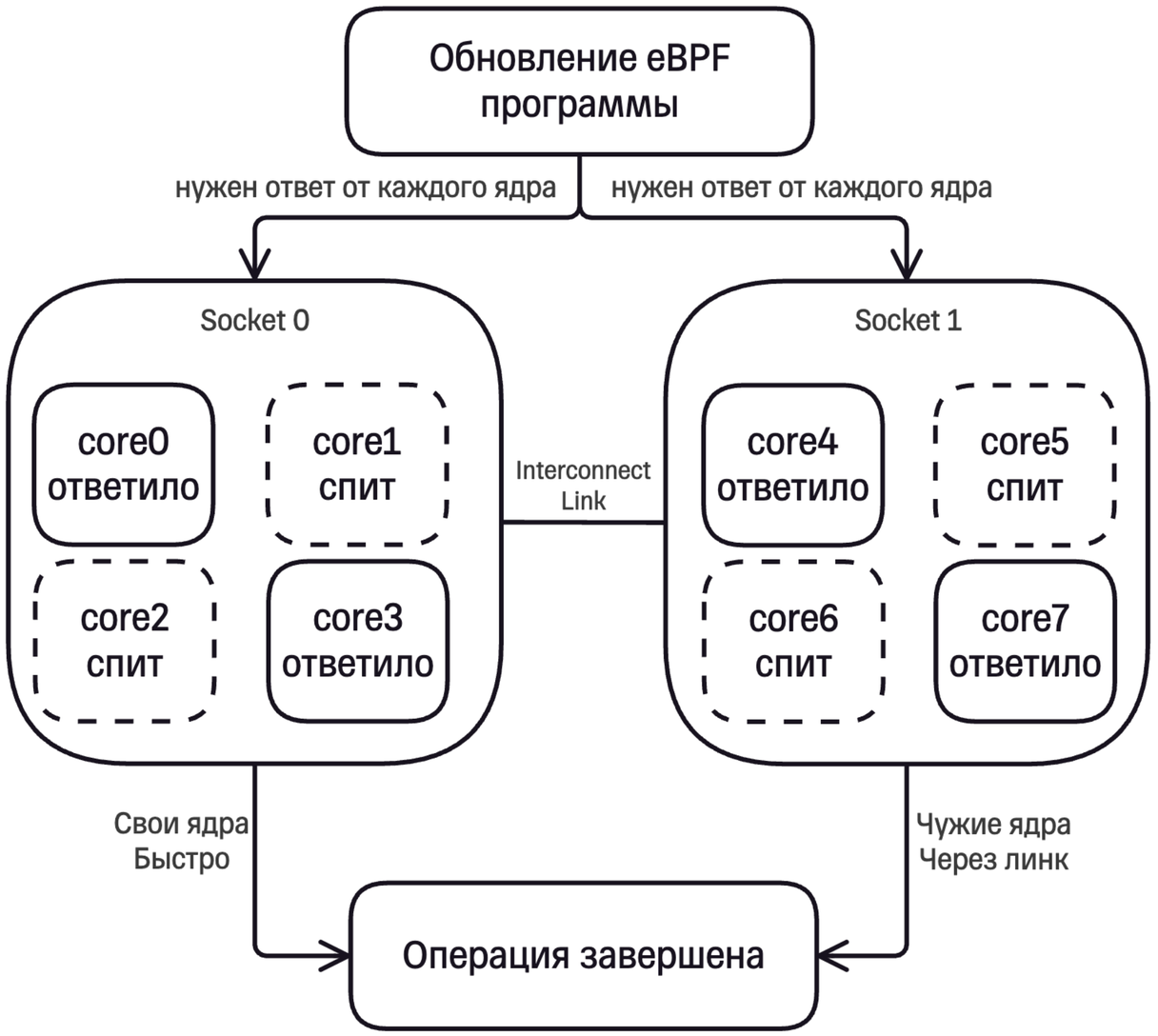
Interconnect PM

- Управление – CPU, данные – NVSwitch Fabric
- На системах с двумя сокетами есть Interconnect Link
- При спящем линке тратим время на его пробуждение



Бонус
Cgroups v2

- При обновлении eBPF каждое ядро должно подтвердить, что eBPF больше не используется
- Спящий линк мог удлинять время обновления eBPF



Итоговые выводы

Запустили два высокопроизводительных GPU кластера

Улучшили надежность инструментов на основе opensource-моделей

Разобрали весь NVIDIA stack

- Сформировали набор тестов и референсных значений
- Далеко не всегда root cause оказывается действительным root cause

Дополнительные материалы



Спасибо за внимание!

Ваши вопросы



No-code
Python
JavaScript
Node.js
Swift
Mobile App
Development
Innovation
Golang
Java
Analysis
Planning

