Odin

Живая миграция контейнеров:

плюсы, минусы, подводные камни

Павел Емельянов

Калуга, 2015

Про что доклад

- Почему надо мигрировать контейнеры
- Почему не надо мигрировать контейнеры
- Насколько сложно мигрировать контейнеры

Миграция в общих чретах

- Сохранить состояние
- Скопировать состояние
- Восстановить состояние

Миграция контейнеров

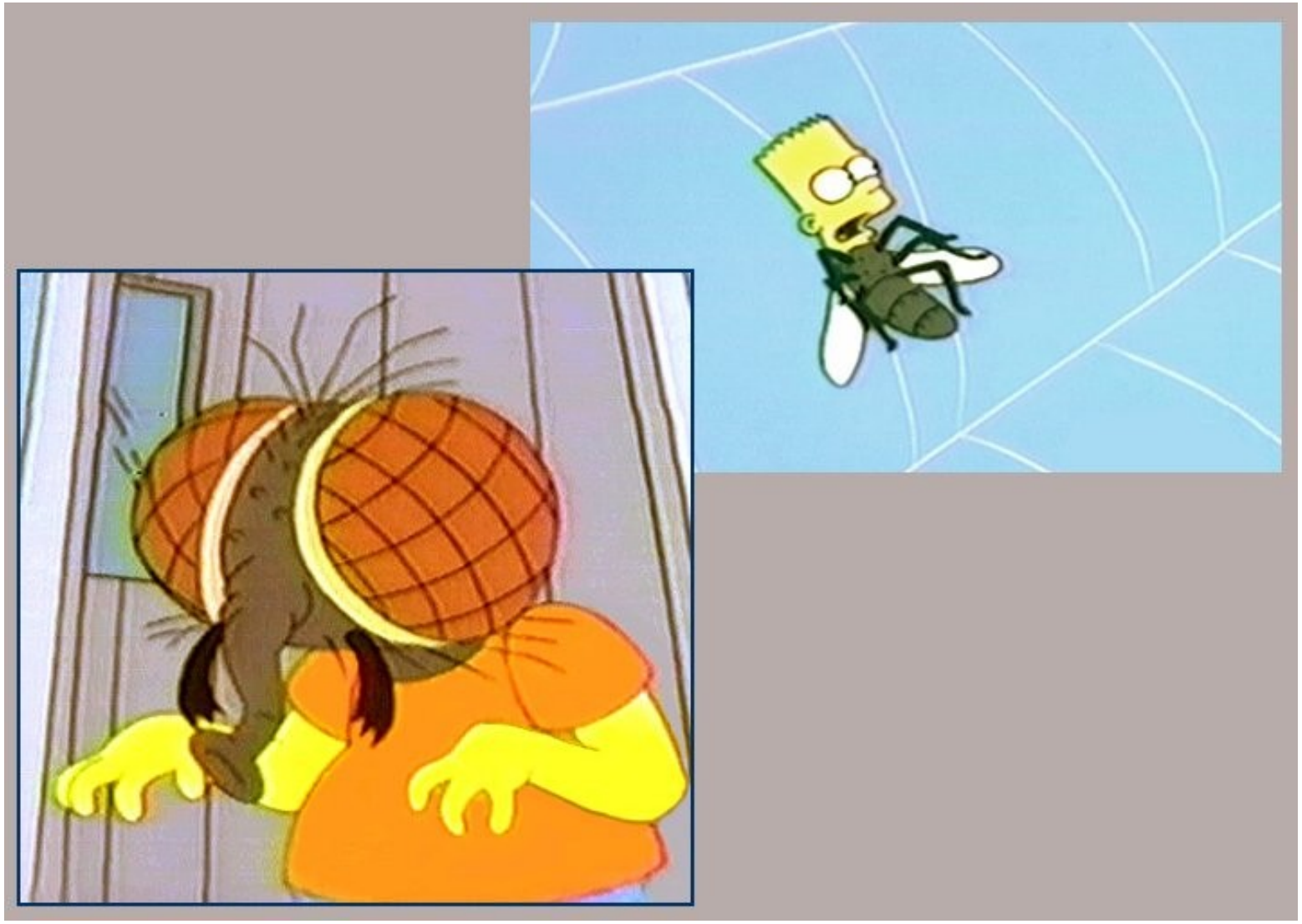
OpenVZ
Virtuozzo Containers

P.Haul

Почему надо мигрировать контейнеры

- Эффектно
- Балансировка нагрузки
- Обновление ядра
  - Можно не мигрировать на самом деле
- Замена оборудования

Почему не надо мигрировать контейнеры



Как не мигрировать контейнеры

- Балансировка сетевого трафика
- Микросервисы
- Crash-driven обновления
- Плановые отключения горячей воды

На самом деле живая миграция

- Пересылку память необходимо исключить из состояния “заморожено”
- Пред-копирование памяти
- Пост-копирование памяти

Живая миграция в деталях

- Пред-копирование: сбор и пересылка памяти (несколько раз)
- Заморозка
- Сохранение состояния
- Копирование состояния
- Восстановление состояния
- Разморозка
- Пост-копирование: подкачка памяти по сети

Подводные камни

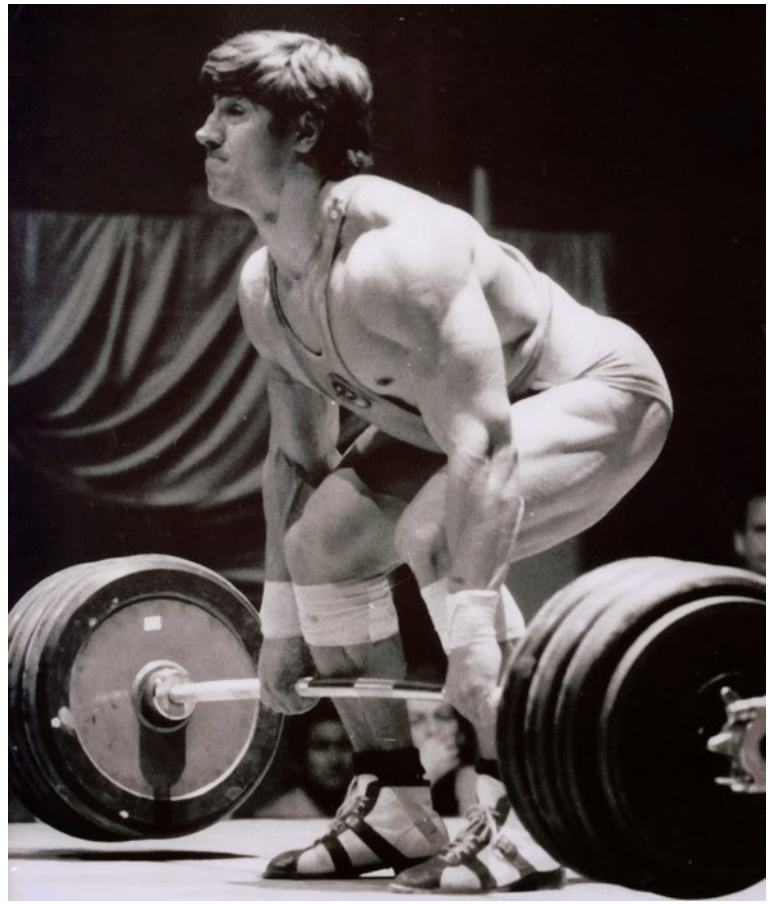
VS

Подводные камни

VS

Подводные камни

VS



Что мигрируем

- VM-ка
  - Окружение (виртуальное железо, paravirt)
  - CPU
  - Память
- Контейнер
  - Окружение (cgroups, namespaces)
  - Процессы и другие животные
  - Память

Сбор и пересылка памяти

- VM-ка
  - Вся память “в руках”
- Контейнер
  - Память размазана по процессам, может быть разделена между ними
  - Поэтому надо сначала собрать процессы (см. ниже)
    - А потом собрать память

Заморозка

- VM-ка
  - Suspend всех процессоров
- Контейнер
  - Пройти по дереву (/proc), переловить процессы и остановить их
  - Freeze cgroup помогает, но надо отдельно восстанавливать иерархию

Сохранение состояния

- VM-ка
  - Состояние железа
    - Дерево, 300K, ~70 объектов
- Контейнер
  - Состояние всех объектов
    - Граф, 160K, ~1000 объектов
    - Не для всех объектов есть адекватный API для чтения

Копирование состояния

- VM-ка
  - Можно читать состояние и сразу передавать
- Контейнер
  - Сложно читать и сразу передавать

Восстановление состояния

- VM-ка
  - Воссоздание памяти, запись состояния в устройства и CPU
- Контейнер
  - В ядре: создание большого количества маленьких объектов
  - В CRIU: то же самое, но с использованием не всегда удобного API
    - Требуется вычисление нетривиальной последовательности действий

Разморозка

- VM-ка
  - Resume
- Контейнер
  - Синхронизация восстановления всех процессов, чтобы не разморозить кого не следует раньше времени
  - SIGCONT по дереву
  - “Оттаять” cgroup

Подкачка памяти по сети

- UserfaultFD от Андреа Арканджели
- VM-ка
  - Merged into 4.2
- Контейнер
  - Несовместная работа монитора и процесса – надо доделывать uffd

Реализация

- http://criu.org
- criu@openvz.org
- +CriuOrg
- @__criu__
- Github: xemul/criu

Реализация

- P.Haul (Пихль)
  - http://criu.org/P.Haul
  - Миграция с помощью CRIU

Odin

Всё.

xemul@openvz.org