Parallels™

Checkpoint-restore in userspace. Готово или нет?

Павел Емельянов, Кир Колышкин

Yet another conference

1 октября 2012

Что такое C/R и зачем оно нужно?

C/R – это возможность сохранить полное состояние приложения и восстановить из него это приложение потом.

Зачем:

- Живая миграция
- Ускорение старта тяжелых приложений
- Обновление ядра “без перезагрузки”
- Снимки состояний рабочей среды
- Балансирование нагрузки в распределенной вычислительной системе
- ...

И что, можно все эти замечательные вещи делать уже сейчас?

Да!

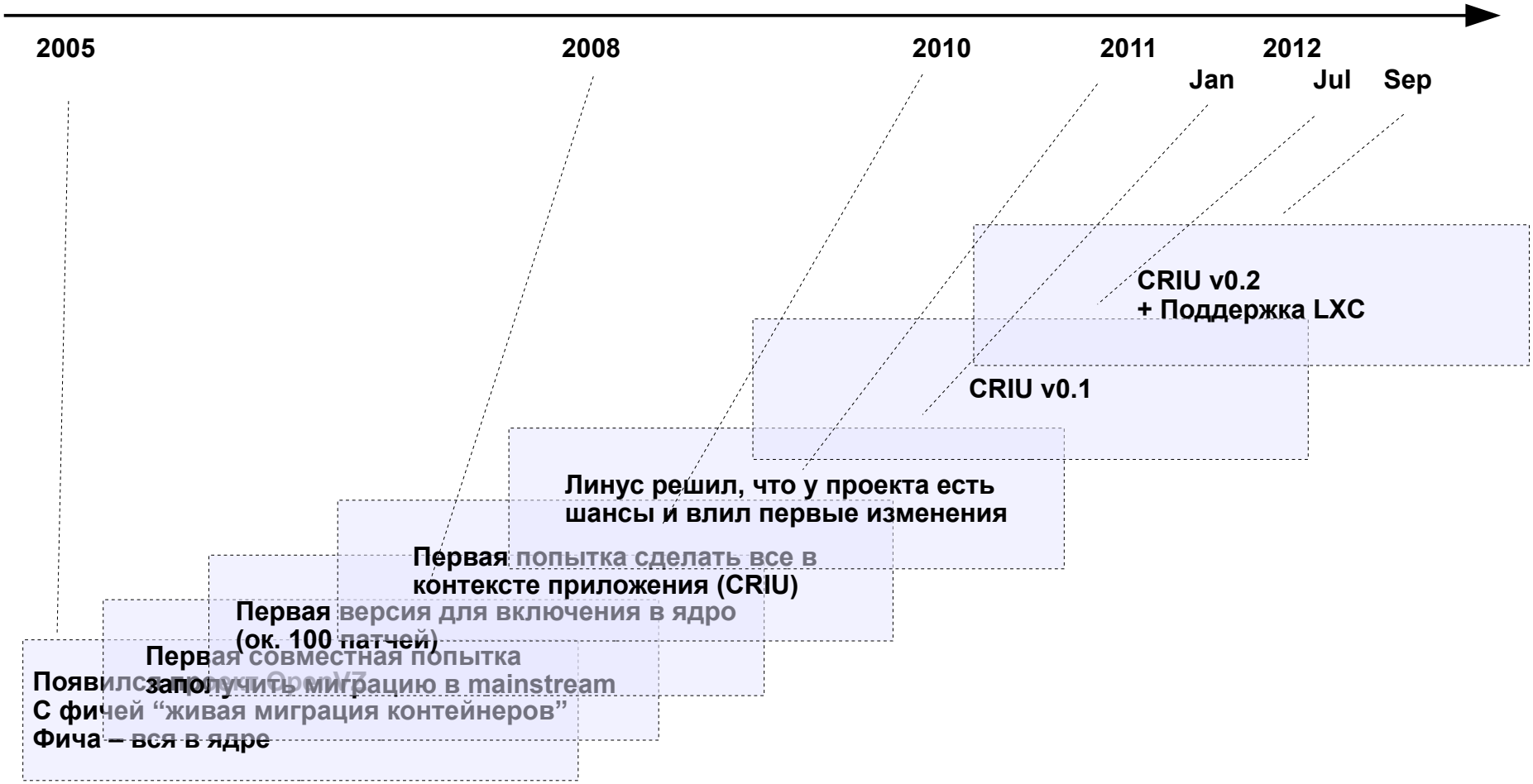
Почти.

Мы близки к этому как никогда!

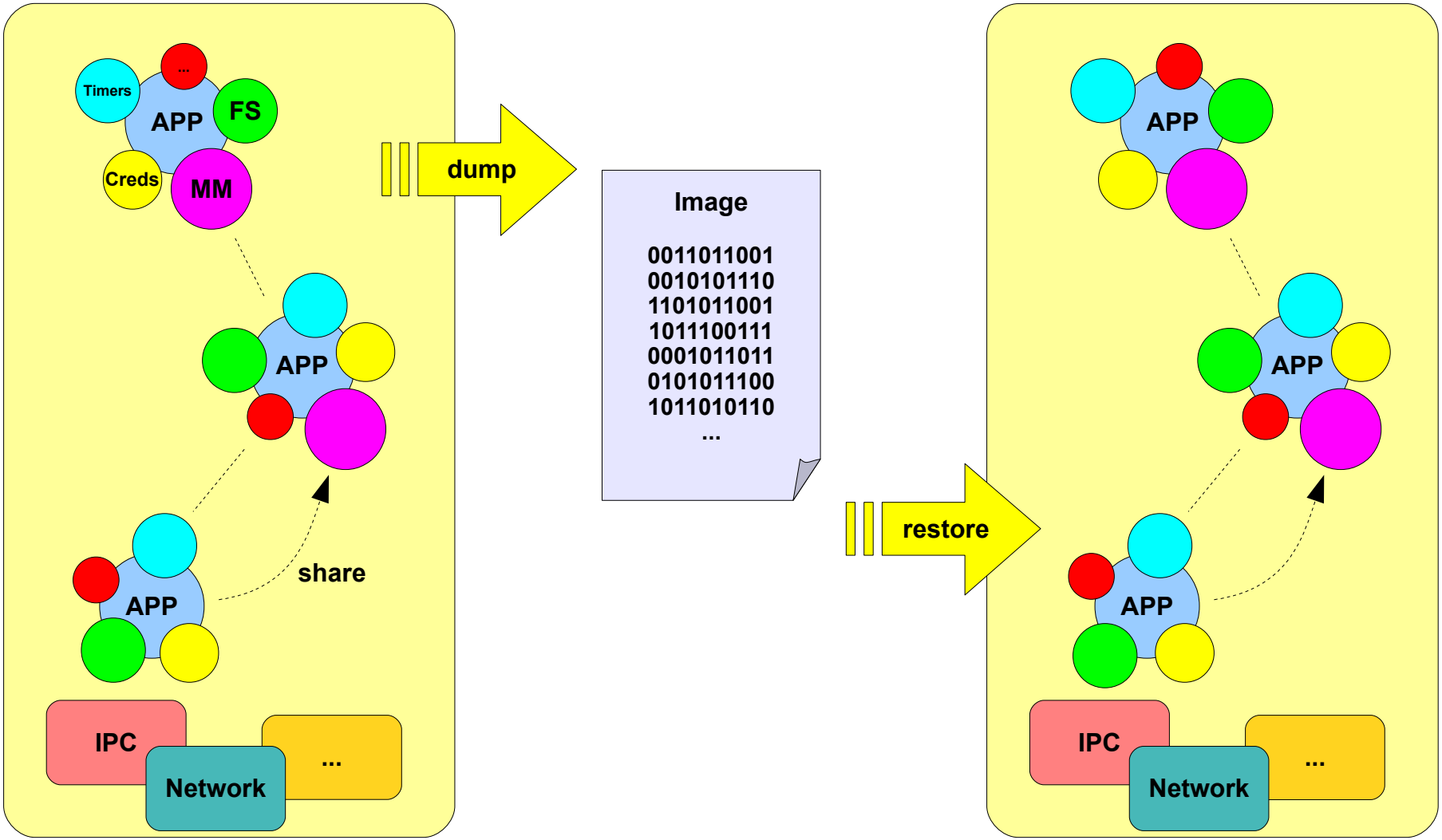
Из доклада вы узнаете:

- Как же все-таки это сделать?
- Долго ли еще ждать?
- Что было сделано, чтобы это стало возможным?

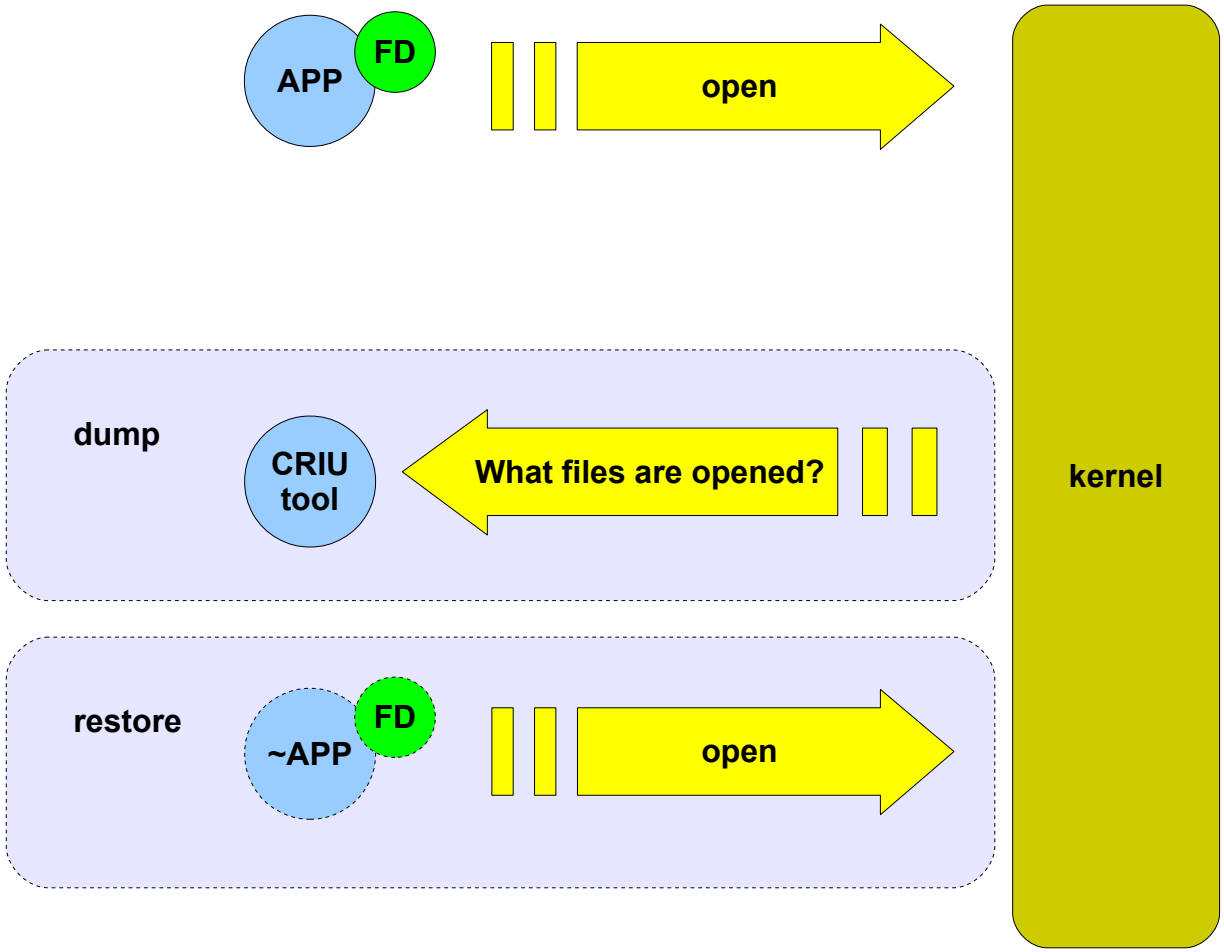
Исторический экскурс



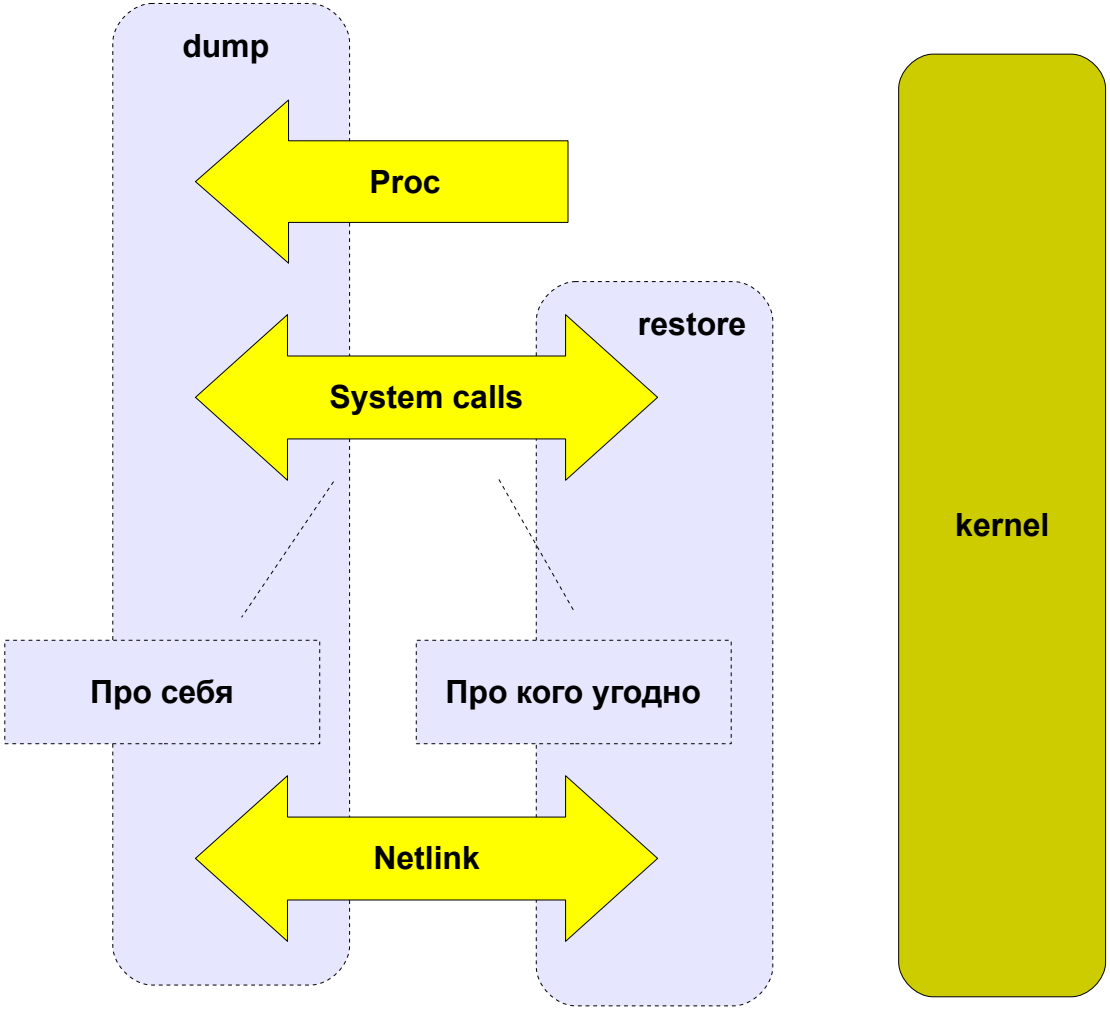
Заветная цель проекта



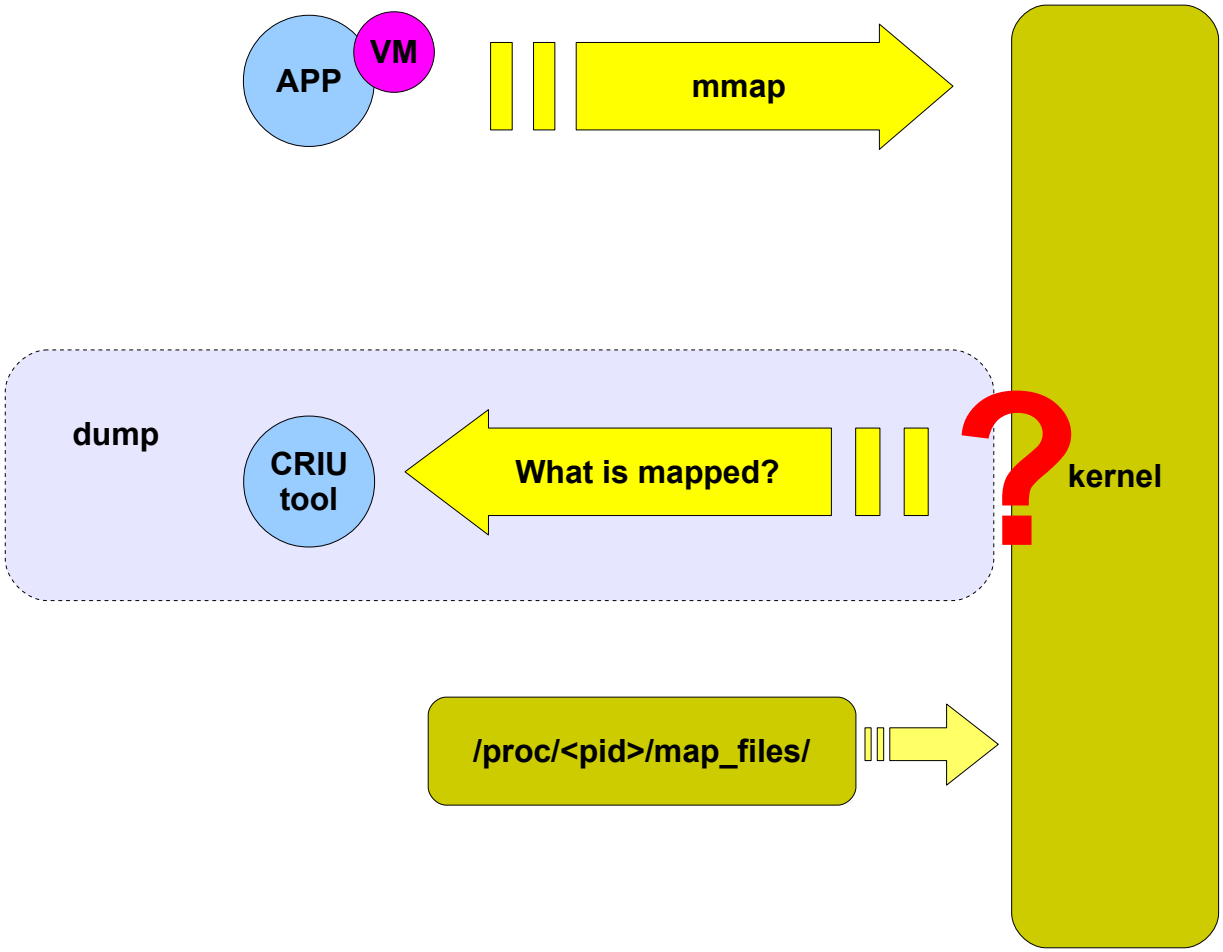
Основополагающий принцип



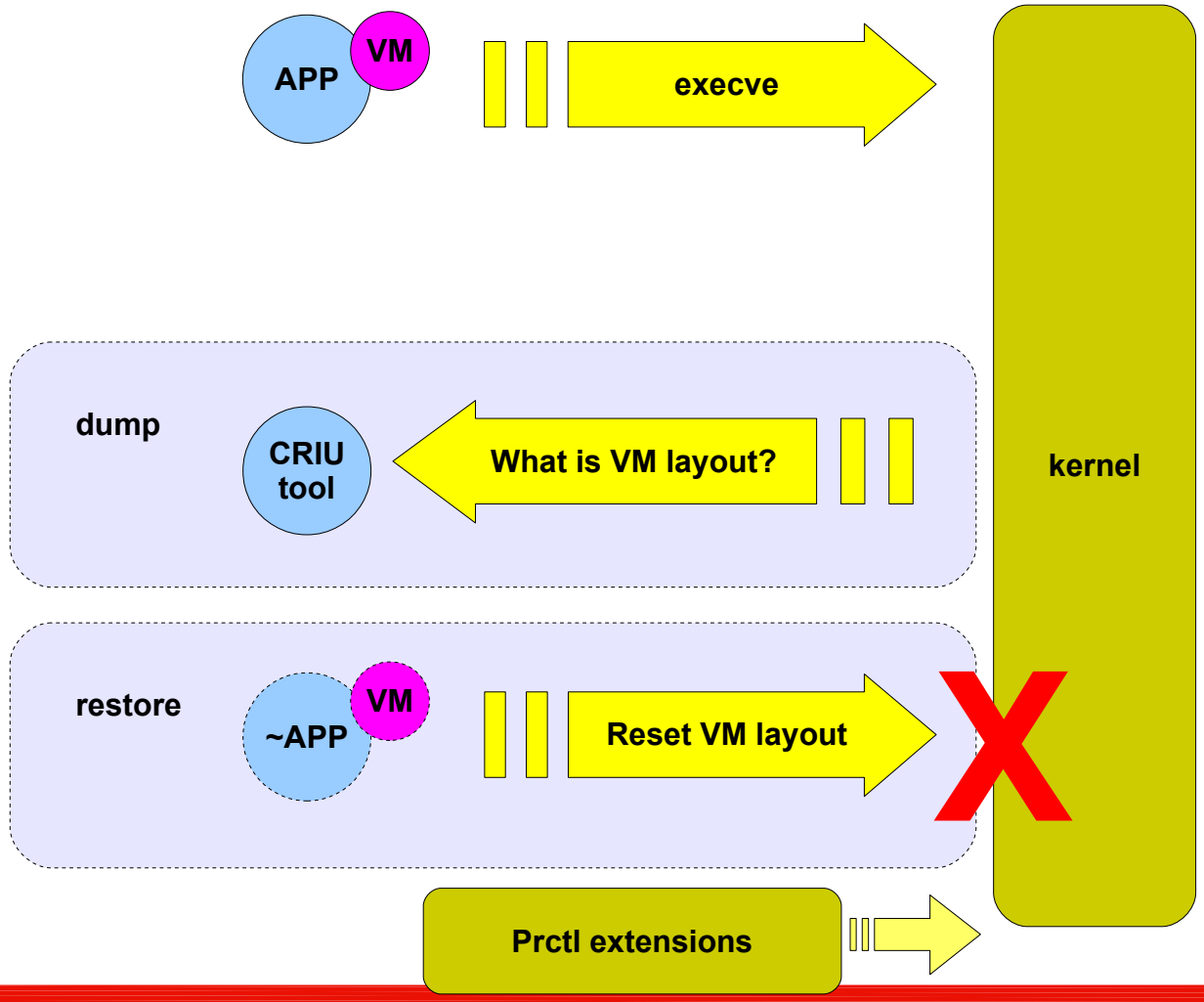
Какие интерфейсы есть у ядра



Как растет CRIU



Как растет CRIU (продолж.)

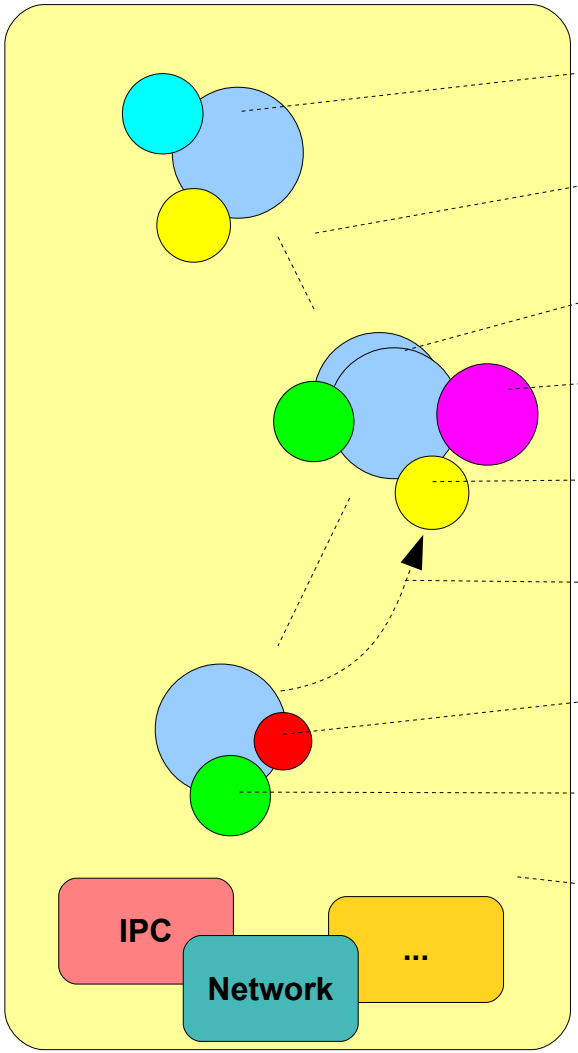


Видение проекта Линусом

... this is a project by various mad Russians to perform c/r mainly from userspace, with various oddball helper code added into the kernel where the need is demonstrated.

So rather than some large central lump of code, what we have is little bits and pieces popping up in various places which either expose something new or which permit something which is normally kernel-private to be modified...

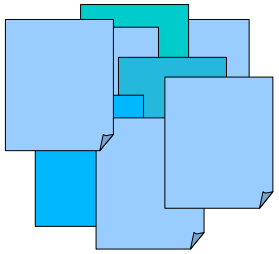
Что уже умеет CRIU



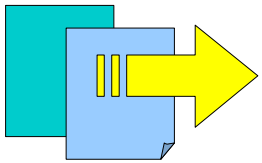
- X86_64 архитектура
- Произвольное поддерево процессов
- Многопоточные приложения
- *Любые* участки виртуальной памяти
- Открытые файлы (удаленные тоже)
- Разделяемые файловые дескрипторы
- Живущие TCP соединения
- Сокеты UNIX
- Контейнеры LXC
- **Kernel V3.5**

Что изменилось в ядре

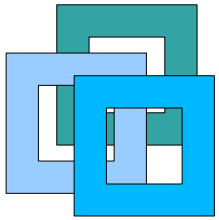
~100 патчей влито



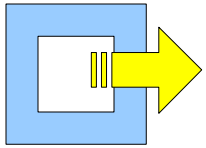
~15 патчей на подходе



9 новых фич
(1 C/R-only)



2 фичи скоро будут



Обзор новых фич в ядре (core)

- Parasite code injection
  - Read task states, that are currently retrieved by a task only about himself
- Proc map_files directory
  - Find out what *exact* file is mapped
  - Mappings sharing info
- The kcmp system call
  - Helps checking which kernel objects are shared between processes

Обзор новых фич в ядре (net)

- TCP repair mode
  - Read intimate state of a TCP connection
    and reconstructs it from scratch on a freshly created socket
- Sockets information dumping via netlink (sock_diag)
  - Extendable sockets state retrieving engine
- Virtual net devices indices
  - Allows to restore network devices in a namespace
- Socket peeking offset
  - Allows peeking sockets queues
    (reading without removing data from queue)

Обзор новых фич в ядре (misc)

- Last-pid sysctl
  - Restore task with desired PID value
- A bunch of prctl extensions
  - Set various private stuff on task/mm objects (c/r-only feature)

Как этим пользоваться

```
# ps axf
...
42   my_app
43     `- my_app_kid
...
# crtools dump -t 42 -D dump/
# ls dump/
pstree.img   core-42.img   fdinfo-42.img   reg-files.img
...
# crtools show -f dump/pstree.img
Pid: 42  ppid: 0
Pid: 43  ppid: 42
# crtools restore -t 42 -D dump/ -d
# ps axf
...
42   my_app
42     `- my_app_kid
...
```

Как тестируется

- ZDTM – набор атомарных тестов
- Real-life приложения
  - Apache
  - MySQL
  - make & gcc
  - tar & gzip
  - sshd с открытыми удалёнными сессиями
  - screen с внутренностями
  - Java
  - VNC сервер + Xscreensaver с клиентской сессией

Главные планы на ближайшее будущее

- Стабилизировать
- Довливать патчи, которые еще не влиты
  Чтобы *все* работало на mainstream linux
- Интегрировать CRIU в OpenVZ и LXC
- Автоматизировать живую миграцию
- Оптимизировать миграцию памяти приложений

Ресурсы проекта

http://criu.org – новости и документация

http://git.criu.org – git с исходниками утилиты

https://github.com/cyrillos/linux-2.6/ – ядро с еще не влитыми патчами

criu@openvz.org – список рассылки

Павел Емельянов
xemul@parallels.com

Кир Колышкин
kir@openvz.org