Linux Control Groups
(краткое введение)

Темы

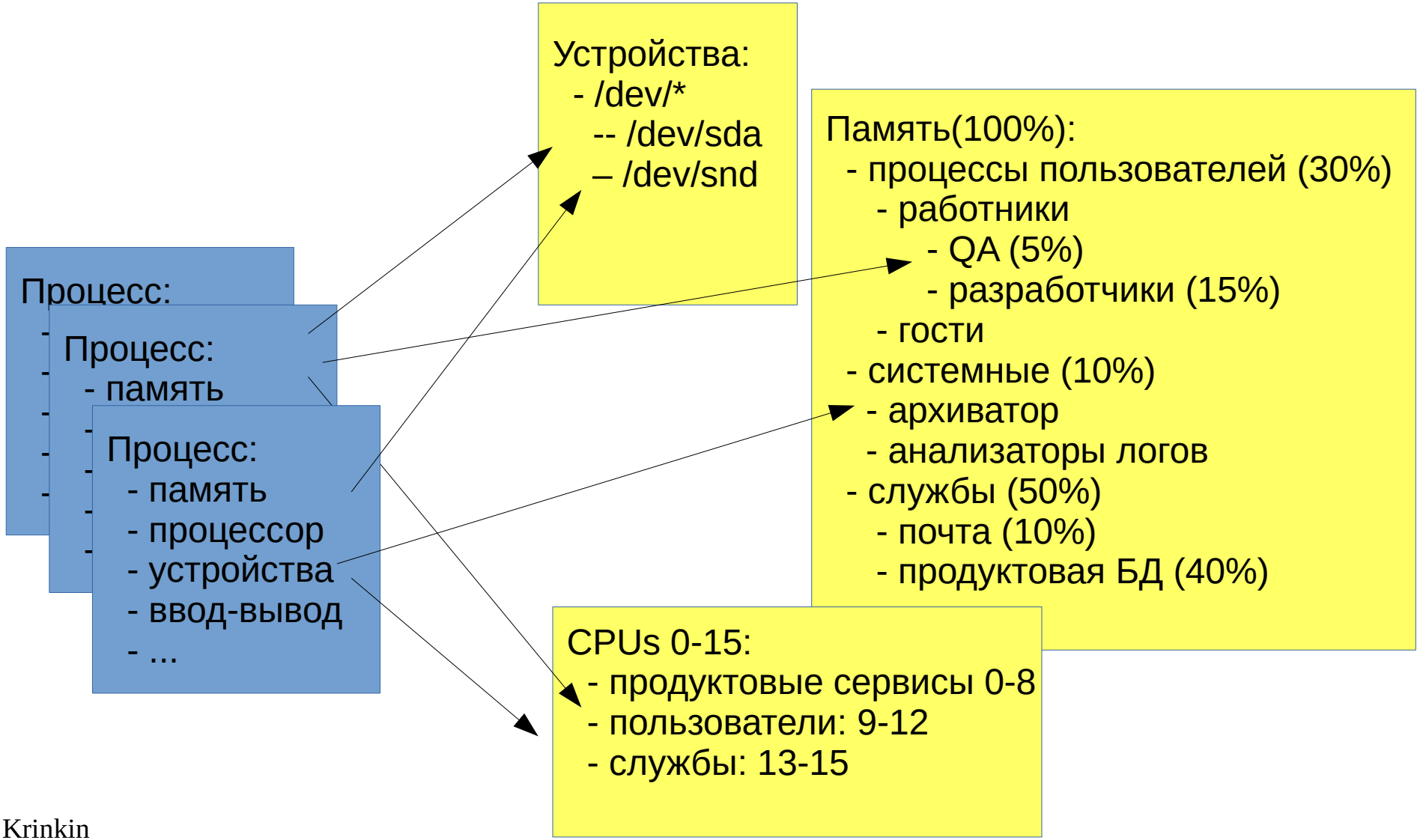
- Подсистемы Linux (немного истории)
- Иерархия ресурсов и контрольные группы
- Обзор контроллеров
- Примеры

Подсистема и контрольная группа

- **Подсистема (subsystem)** – модуль предоставляющий возможности для группировки и управления (ограничение, подсчет использования,...) определенными ресурсами процессов.

- **Контрольная группа (Control Group, cgroup)** – множество параметров, ассоциированное с одной или более подсистем.

Иерархии ресурсов



Подсистемы = контроллеры

- blkio – управление вводом выводом блочных устройств
- cpu – управление доступом к процессору
- cpuacc – отчеты по использованию процессора
- cpuset – приязка к процессорам и банкам памяти
- devices – доступ к устройствам
- freezer – останов/возобновление работы группы
- memory – ограничения и учет использования памяти
- net_cls – маркировка пакетов для контроллера трафика

Файлы cgroup

- /proc/cgroups
- /proc/self/cgroup – группы процесса
- /sys/fs/cgroup/ (/cgroup) – корень иерархии
- */tasks – PIDs участников группы
- */cgroup.procs – список thread groups
- */notify_on_release – флаг вызова агента разрушения (по-умолчанию 0)
- */release_agent – путь к агенту разрушения

Подсистемы и точки монтирования

```
kkv@thinkpad:~$ lssubsys -am
cpuset /sys/fs/cgroup/cpuset
cpu /sys/fs/cgroup/cpu
cpuacct /sys/fs/cgroup/cpuacct
memory /sys/fs/cgroup/memory
devices /sys/fs/cgroup/devices
freezer /sys/fs/cgroup/freezer
blkio /sys/fs/cgroup/blkio
perf_event /sys/fs/cgroup/perf_event
hugetlb /sys/fs/cgroup/hugetlb
kkv@thinkpad:~$
```

```
kkv@thinkpad:~$ lscgroup
cpuset:/
cpuset:/user
cpuset:/user/1000.user
cpu:/
cpu:/user
cpu:/user/1000.user
cpuacct:/
cpuacct:/user
...
```

cgcreate: создание группы

- **cgcreate**

  -t uid:gid – пользователи получающие права на перемещение заданий в(из) группу

  -a uid:gid – пользователи получающие права на управление параметрами группы

  -g список подсистем (контроллеров):путь

- Пример

```
cgcreate -t kkv:kkv -g memory,cpu:/mycgroup
```

cgdelete: удаление группы

- **cgdelete**

  -g список подсистем (контроллеров):путь

- Пример

```
cgdelete -g memory,cpu:/mycgroup
```

(*) при удалении группы, входящие в нее задачи перемещаются в родительскую группу

Перемещение процессов в группу

- **cgclassify**

  -g список подсистем (контроллеров):путь

  PID [PID PID …]

Примеры:

```
$cgclassify -g cpu:/mycgroup 6433 3662
$echo 6433 >/sys/fs/cgroup/cpu/mucgroup/tasks
```

Выполнение процесса в группе

- **cgexec**

  -g список подсистем (контроллеров):путь имя_приложения

- Пример

```
cgexec -g memory:/mycgroup google-chrome
```

Группы процесса

```
kkv@thinkpad:$ cat /proc/$$/cgroup
11:name=systemd:/user/1000.user/c2.session
10:hugetlb:/user/1000.user/c2.session
9:perf_event:/user/1000.user/c2.session
8:blkio:/user/1000.user/c2.session
7:freezer:/user/1000.user/c2.session
6:devices:/user/1000.user/c2.session
5:memory:/seminar/memory
4:cpuacct:/user/1000.user/c2.session
3:cpu:/user/1000.user/c2.session
2:cpuset:/seminar/cpuset
```

cgget – доступ к параметрам

- **cgget**

  -r параметр группа

  Пример:

```
$cgget -r cpu.shares /mycgroup
>/mycgroup:
>cpu.shares: 1024

cat /sys/fs/cgroup/cpu/mycgroup/cpu.shares
>1024
```

cgset – установка параметров

- **cgset**

  -r параметр=значение

Пример:

```
$cgset -r cpu.shares=8 /mycgroup
```

```
$echo 9 >/sys/fs/cgroup/cpu/mycgroup/cpu.shares
```

Контроллер cpuset

- Назначение: управление привязкой процессоров и памяти к процессам
- ←/→ cpuset. cpus – список привязанных процессоров
- ←/→ cpuset. mems – список привязанных процессоров
- ←/→ cpuset. cpu_exclusive – флаг эксклюзивного использования процессора группой
- ←/→ cpuset.sched_load_balance + cpuset.sched_relax_domain_level – управление балансировкой нагрузки в группе

try: cat /proc/self/status

cpuset.sched_relax_domain_level

- -1 – не менять внешние правила
- 0 – периодическая балансировка
- 1 – немедленная между потоками одного ядра
- 2 – немедленная между ядрами пакета
- 3 – немедленная в рамках узла
- 4 – немедленная между процессорами на NUMA системе
- 5 – немедленная по всей системе

try: cat /proc/self/status

Контроллер cpu

- Назначение: управление распределеним нагрузки на процессоры

  cpu.shares – доля использования процессора по отношению к другим группам

  cpu.rt_runtime_us – максимальноый период монопольного использования процессора в микросекундах

  cpu.rt_period_us – максимальное время ожидания процессора группой

Контроллер cpuacct

- Назначение: сбор статистики по использованию процессора

  cpuacct.stat – число циклов процессора

  cpuacct.usage – суммарное время

  cpuacct.usage_percpu – число циклов процессора, включая задания подгрупп

Контроллер devices

- Назначение: управление доступом к устройствам из группы
  - → devices.allow – устройства доступные группе
  - → devices.deny – запрещенные устройства в группе
  - ← devices.list – просмотр устройств в whitelist

  Примеры:

```
echo 'c 1:3 mr' > /sys/fs/cgroup/1/devices.allow
echo a > /sys/fs/cgroup/1/devices.deny
echo a > /sys/fs/cgroup/1/devices.allow
```

https://www.kernel.org/doc/Documentation/devices.txt

Контроллер freezer

- Назначение: заморозка/разморозка исполнения группы процессов

  ←/→freezer.state – состояние заморозки

  - FROZEN — задания приостановлены
  - FREEZING – в стадии приостановки (включая группы-потомки)
  - THAWED – возобновление работы

- ←freezer.self_freezing собственное состояние заморозки
- ←freezer.parent_freezing родительское состояние состояние заморозки

Контроллер memory

- Назначение: управление и мониторинг использования памяти
  - ←memory.stat – получение статистики по использованию памяти
    - total_ – текущая группа и подгруппы
  - ←memory.[memsw.]usage_in_bytes – используемая память в байтах (в подкачке)
  - ←/→memory.[memsw.]limit_in_bytes
  - ←memory.[memsw.].failcnt – счетчик числа достижений лимита памяти
  - ←/→memory.oom_control — флаг разрешения OOM-killer
  - (*) ←/→memory.soft_limit_in_bytes — флаг разрешения OOM-killer

см https://www.kernel.org/doc/Documentation/cgroups/memory.txt

Контроллер blkio

- → blkio.weight – [100-1000], относительный вес ввода вывода в группе
- → blkio.weight – [100-1000], относительный вес ввода вывода в группе для конкретного устройства
- ← blkio.time – время доступа ввода-вывода в группе
- ← blkio.sectors – количество перемещенных между устройствами секторов в группе
- ← blkio.io_service_bytes – количество перемещенных между устройствами байт в группе
- ← blkio.io_service_time – время между выдачей запроса и его завершением
- ← blkio.io_queued – число запросов в очереди ввода вывода группы

Для чтения

- RHEL System resources management guide
- https://www.kernel.org/doc/Documentation/cgroups/cgroups.txt
- https://www.kernel.org/doc/Documentation/cgroups/memory.txt

Примеры для изучения

- Элементы:
  - создание группы
  - перемещение и запуск процесса
  - изменение параметров группы
- cpuset
  - привязка cpu
- memory
  - установка лимита памяти и oom-killer
- cpu
  - разделение веса использования процессора между группами
- freezer
  - заморозка/разморозка группы