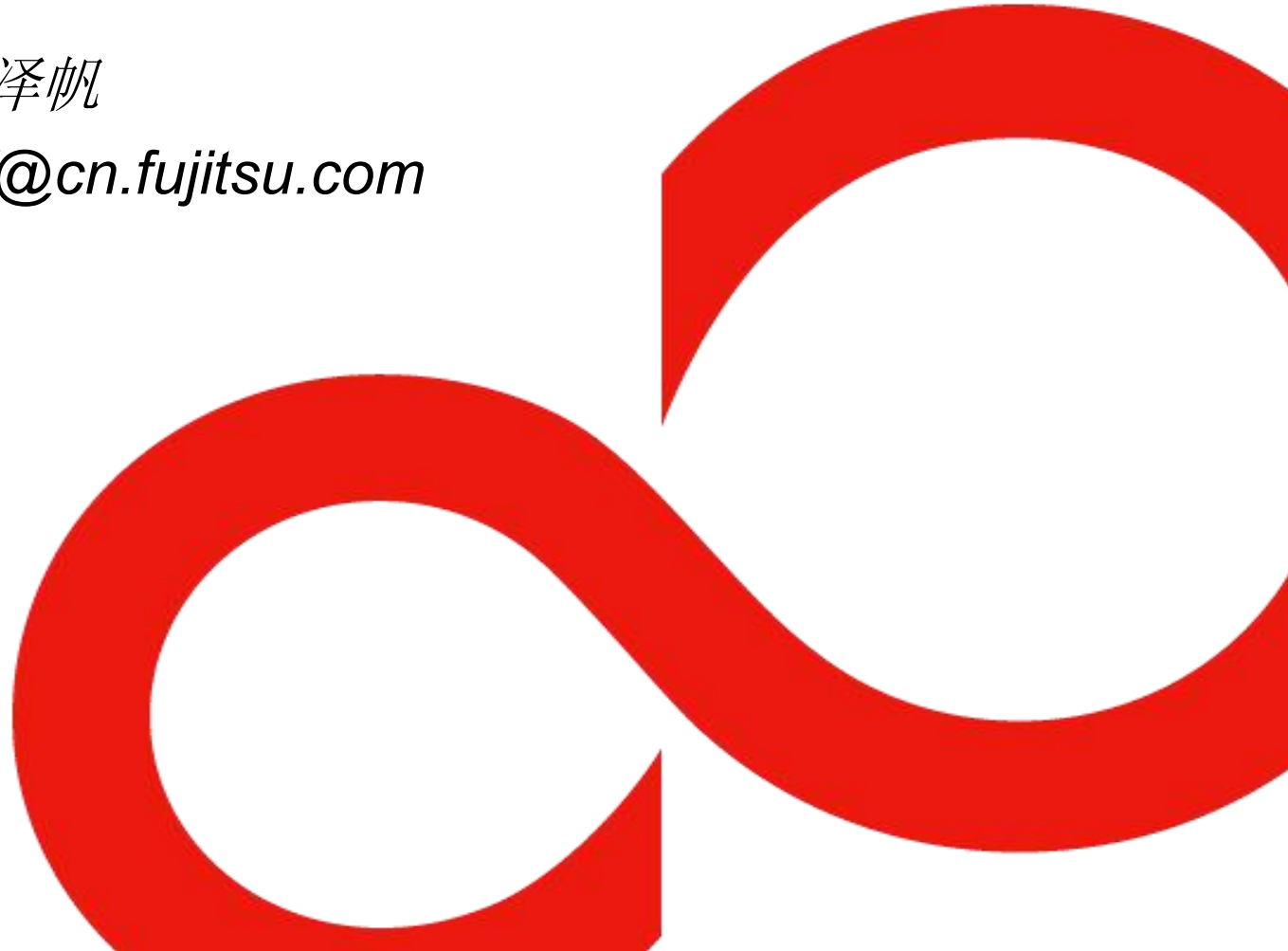
THE POSSIBILITIES ARE INFINITE FUJITSU

•资源管理—cgroup及其子系统

李泽帆

lizf@cn.fujitsu.com



资源管理—cgroup及其子系统

What is cgroup

The interface and implementation of cgroup

Problems with cgroup

Introduction of cgroup subsystems

What is cgroup

per-process资源管理：rlimit

CKRM：Class-based Kernel Resource Management

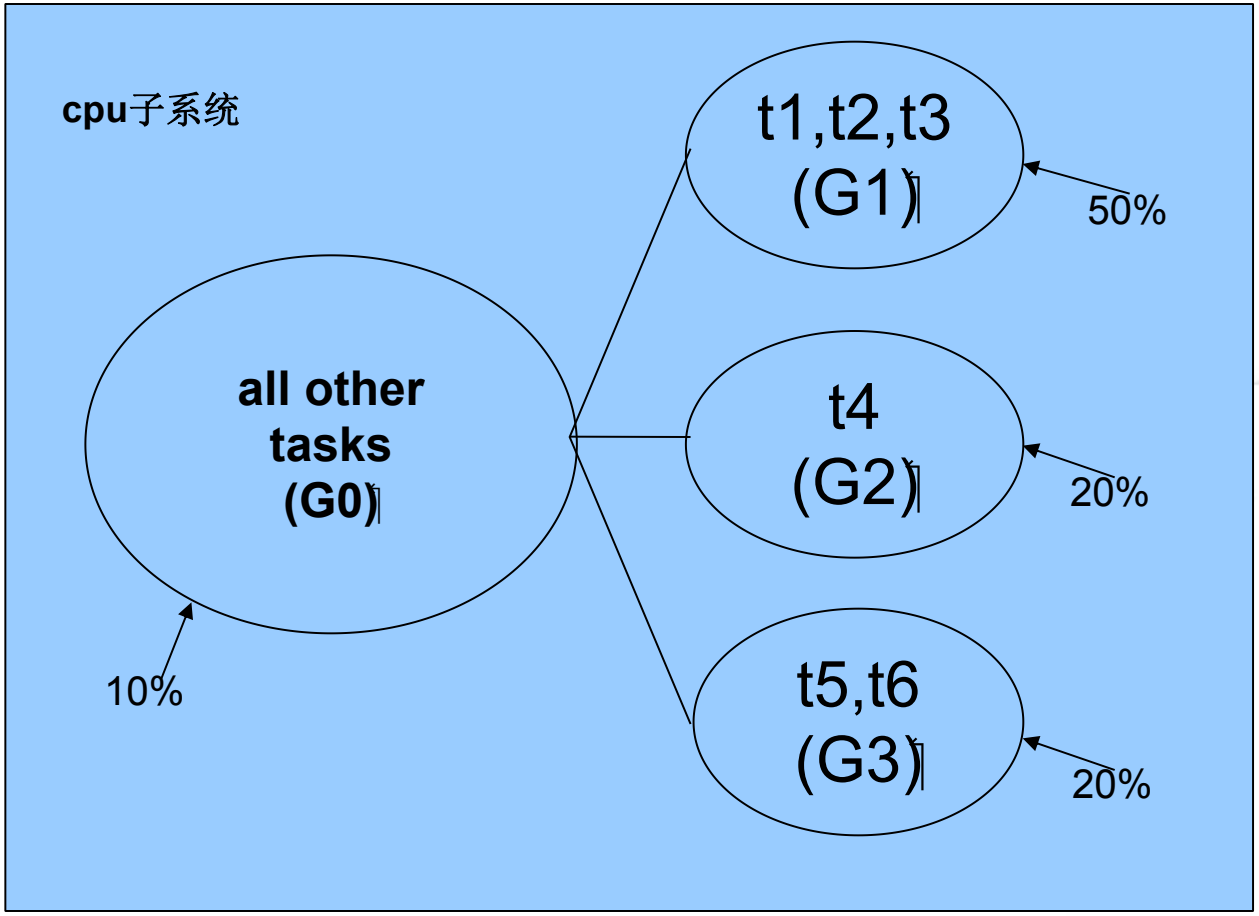
BeanCounters

cgroup(control group)：对系统进程进行分组

cgroup subsystem：使用cgroup的分组机制，对一组进程就某种系统资源实现资源管理。

What is cgroup

图例



cgroup的接口与实现

What is cgroup

The interface and implementation of cgroup

Problems with cgroup

Introduction of cgroup subsystems

cgroup的接口与实现

用户接口：**pseudo-filesystem**

cgroup <-> cgroup subsys

~

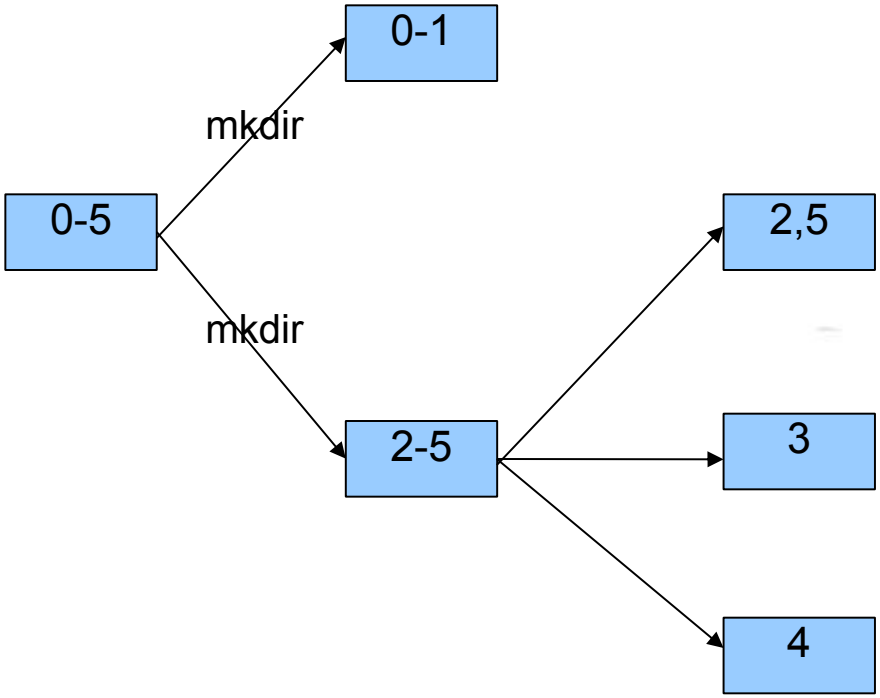
vfs <-> filesystem (ext4, btrfs, reiserfs ...)

mount -t cgroup -o cpuset, memory, devices xxx /mnt

```
[root@localhost /]# mount -t cgroup -o cpu,memory,devices xxx /cgroup/
[root@localhost /]# ls /cgroup/
cpu.rt_period_us   devices.deny        memory.limit_in_bytes      notify_on_release
cpu.rt_runtime_us  devices.list        memory.max_usage_in_bytes  release_agent
cpu.shares         memory.failcnt      memory.stat                tasks
devices.allow      memory.force_empty  memory.usage_in_bytes
[root@localhost /]# mkdir /cgroup/Group1
[root@localhost /]# echo 2596 > /cgroup/Group1/tasks
[root@localhost /]# cat /cgroup/Group1/tasks
2596
[root@localhost /]# echo 100M > /cgroup/Group1/memory.limit_in_bytes
```

cgroup的接口与实现

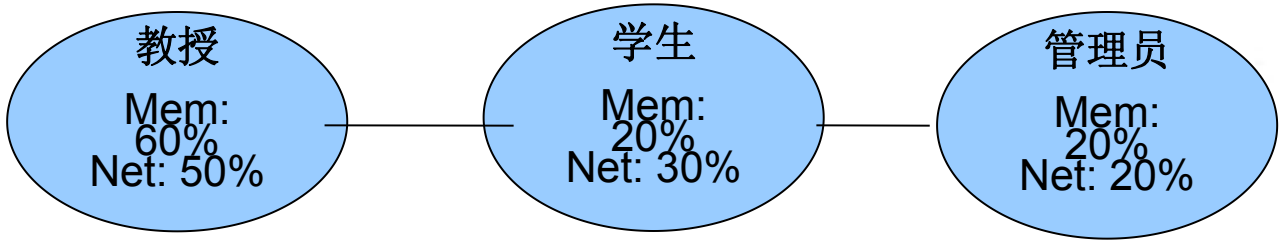
cgroup的树结构
(以cpuset为例)



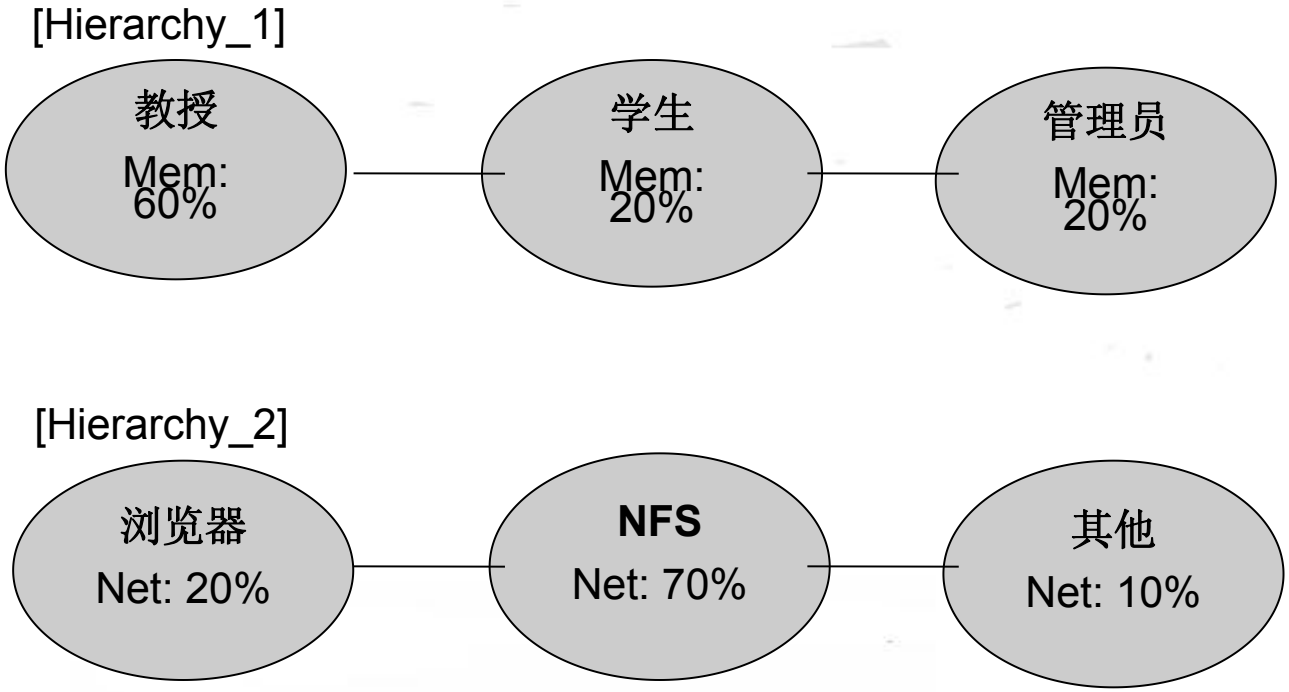
cgroup的接口与实现

Single or multiple hierarchies

Single：



Multiple：



cgroup的接口与实现

进程移动

fork：do_fork() -> copy_process() -> cgroup_fork()

-> child->cgroups = current->cgroups;

exit：do_exit() -> cgroup_exit()

-> tsk->cgroups = &init_css_set;

attach：cgroup_tasks_write()

-> attach_task_by_pid(new_cgrp, pid)

-> can_attach(tsk, new_cgrp) --> attach()

cgroup的接口与实现

cgroup子系统需要实现的调用接口用户接口

create()　　- 创建cgroup时调用

destroy()　　- 删除cgroup时调用

populate()　　- 生成subsys.xxx控制文件

can_attach()　- 进程可否移动

attach()　　- 移动进程

fork()　　- 新进程fork时调用

exit()　　- 进程结束时调用

cgroup的问题

What is cgroup

The interface and implementation of cgroup

Problems with cgroup

Introduction of cgroup subsystems

cgroup的问题

gap between can_attach() and attach()

```
cgroup_lock();

...

for_each_subsys(root, ss) {
.       if (ss->can_attach) {
.       .       retval = ss->can_attach(ss, cgrp, tsk);
.       .       if (retval)
.       .       .       return retval;
.       }
}

...

for_each_subsys(root, ss) {
.       if (ss->attach)
.       .       ss->attach(ss, cgrp, oldcgrp, tsk);
}

cgroup_unlock();
```

cgroup的问题

“procs” control file

当前：一次只能移动一个进程

echo $pid > /cgroup/sub/tasks

问题1

问题2：名字冲突

procs or cgroup.procs

cgroup的问题

Rules for tasks attaching

daemon using Process Event Connector

wrappers around binaries

script running at ssh logins

kernel-side rule engine

cgroup子系统介绍

What is cgroup

The interface and implementation of cgroup

Problems with cgroup

Introduction of cgroup subsystems

cgroup子系统介绍

cpuset

为一组进程分配一组cpu和内存结点

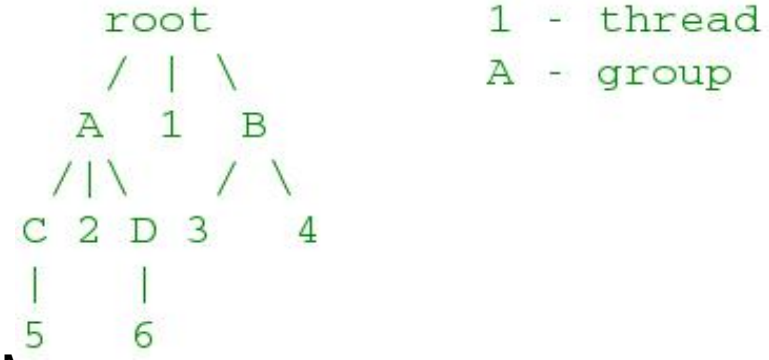
cpuset中的平衡负载(sched load balance)

……

cgroup子系统介绍

cpu – Group CPU Scheduler

一个group有一个cpu.shares，表示该group的cpu时间的权重



SMP fairness的问题

A: 5474(t1)　B: 5475(t2), 5476(t3)　A.shares = B.shares = 1024　nr_cpus = 2

期望值 – t1 : t2 : t3 = 1024 : 1024/2 : 1024/2 = 2 : 1 : 1 = 100% : 50% : 50%

实际值 – t1 : t2 : t3 = 66.6% : 100% : 33.3%

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
5475	root	20	0	5168	436	372	R	99	0.0	0:51.64	cat
5474	root	20	0	5168	436	372	R	68	0.0	0:35.82	cat
5476	root	20	0	5168	432	372	R	32	0.0	0:17.92	cat

cgroup子系统介绍

devices – Device Whitelist Controller

设备白名单：一个组有一份白名单，组内进程只能访问白名单上的设备

设备的权限类型：read, write, mknod

cgroup子系统介绍

memory – Memory Resource Controller

控制的内存：RSS和Page Cache

*RSS: 驻留在RAM中的虚拟内存页面的数目

Swap controller or mem+swap ？

1. mem and swap: xGB mem + yGB swap
2. mem+swap: (x+y)GB (mem+swap)

cgroup子系统介绍

memrlimit – Memory Address Space Controller

Not OOM, but malloc() or mmap() return failure

Better control over how many pages can be swapped out when the cgroup goes over its limit

cgroup子系统介绍

Freezer cgroup

- Freeze all tasks when suspend/hibernate
- Freeze a set of tasks using cgroup freezer subsystem

cgroup子系统介绍

Network Traffic Controller

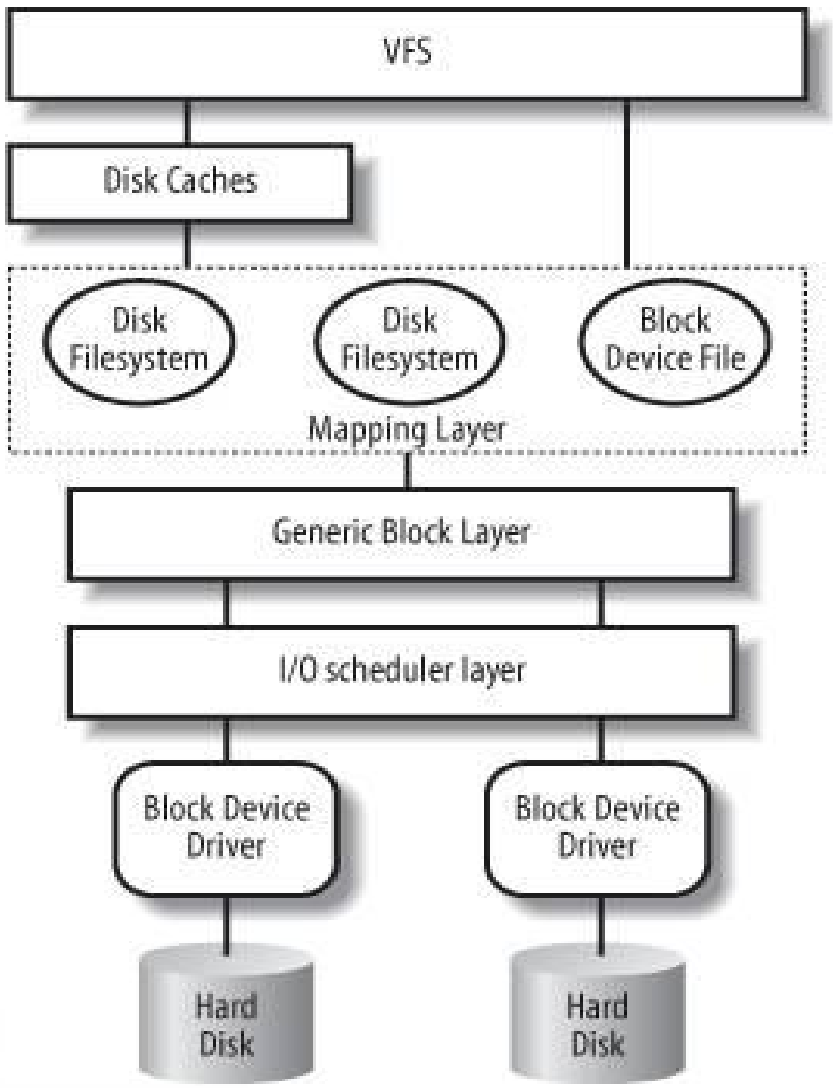
tc_cgroup - Ranjit Manomohan <ranjitm@google.com>

net_cls - Thomas Graf <tgraf@suug.ch>

cgroup子系统介绍

Block I/O Controllerer

- 控制的方式：
  - I/O priorities
  - weight/share
  - bandwidth limiting
- 在哪个layer实现io control：
  - elevator-based io controller
  - block layer io controller



Thanks!