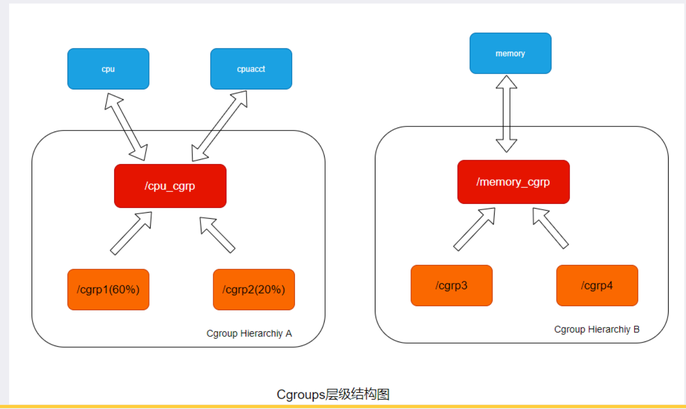
cgroups

- cgroup是linux内核提供的一种机制，可以根据需求把一系列任务及其子任务整合(或分离)到按资源划分等级到不同组内，从而为系统资源管理提供一个统一的框架
- linux伪文件系统
  - 伪文件系统也称为虚拟文件系统
  - 这些文件系统并不是真正的物理存储设备，只在内存中
  - 提供了对系统资源访问的接口，内存、进程、网络
  - /proc文件系统、/dev/shm文件系统
- cgroups特点
  - cgroups的API是一个伪文件系统方式的实现，用户态程序可以通过文件操作实现cgroups的组织管理
  - cgroups的组织管理操作单元可以细粒度到线程级别，另外用户可以创建和销毁cgroups，从而实现资源再分配
  - 所有资源管理功能都以子系统的方式实现，接口统一
  - 子任务创建之处和其父任务处于同一个cgroup控制组
  - 本质上，cgroups是内核附加在程序上的一系列钩子(hook)，通过程序运行时对资源调度触发相应的钩子以达到资源追踪和限制的目的
- cgroups术语
  - task(任务)　任务表示系统中的一个进程或者线程
  - cgroup(控制组)　cgroup中的资源控制都在以cgroup为单位实现。cgroup表示按某种资源控制标准划分而成的任务组，包含一个或者多个子系统。一个任务可以加入某个cgroup也可以迁移到另一个cgroup
  - subsystem(子系统)　资源调度控制器
    - cpu子系统　使用调度程序控制限制cpu的使用
    - cpuacct子系统　统计cgroup中的进程cpu使用报告
    - cpuset子系统　为cgroups中的进程分配单独cpu节点或者内存节点
    - memory子系统　限制进程内存使用量
    - blkio子系统　可以为块设备设定输入/输出限制，比如物理驱动设备(包括磁盘、固态硬盘)
    - devices子系统　限制进程访问某些设备
    - net_cls子系统　可以标记cgroup中进程的网络数据包，使用tc(traffic control)对数据包控制
    - freezer系统　挂起或者恢复cgroup中的进程
    - ns子系统　使不同cgroups下面的进程使用不同的namespace
    - perf_event　使用后使cgroup中的任务可以进行统一的性能测试

```
[root@master cpuacct]# mount -t cgroup
cgroup on /sys/fs/cgroup/systemd type cgroup (rw,nosuid,nodev,noexec,relatime,xattr,release_agent=/usr/lib/systemd/systemd-cgroups-agent,name=systemd)
cgroup on /sys/fs/cgroup/perf_event type cgroup (rw,nosuid,nodev,noexec,relatime,perf_event)
cgroup on /sys/fs/cgroup/cpuset type cgroup (rw,nosuid,nodev,noexec,relatime,cpuset)
cgroup on /sys/fs/cgroup/freezer type cgroup (rw,nosuid,nodev,noexec,relatime,freezer)
cgroup on /sys/fs/cgroup/cpu,cpuacct type cgroup (rw,nosuid,nodev,noexec,relatime,cpuacct,cpu)
cgroup on /sys/fs/cgroup/net_cls,net_prio type cgroup (rw,nosuid,nodev,noexec,relatime,net_prio,net_cls)
cgroup on /sys/fs/cgroup/blkio type cgroup (rw,nosuid,nodev,noexec,relatime,blkio)
cgroup on /sys/fs/cgroup/devices type cgroup (rw,nosuid,nodev,noexec,relatime,devices)
cgroup on /sys/fs/cgroup/hugetlb type cgroup (rw,nosuid,nodev,noexec,relatime,hugetlb)
cgroup on /sys/fs/cgroup/pids type cgroup (rw,nosuid,nodev,noexec,relatime,pids)
cgroup on /sys/fs/cgroup/memory type cgroup (rw,nosuid,nodev,noexec,relatime,memory)
```

  - hierarchy(层级)　层级由一系列cgroup以一个树状结构排列而成，每个层级通过绑定对应的子系统进行资源控制



Cgroups层级结构图

- cgrouo与任务之间关系
  - cgroup与任务之间是多对多的关系，他们之间并不直接关联，而是通过一个中间结构把双向的信息关联起来
  - 每个任务结构体task_struct中都包含一个指针，可以查询到对应的cgroup情况，同时也可以查询到各个子系统情况，这些子系统也包含找到任务的指针