Alibaba Group 阿里巴巴集团

intel IBM IT大咖说 知识分享平台

阿里内部集群的内存超卖

阿里巴巴操作系统团队-陶文苇

系统软件事业部　打造具备全球竞争力、效率最优的系统软件

Catalog

01 混部

02 超卖

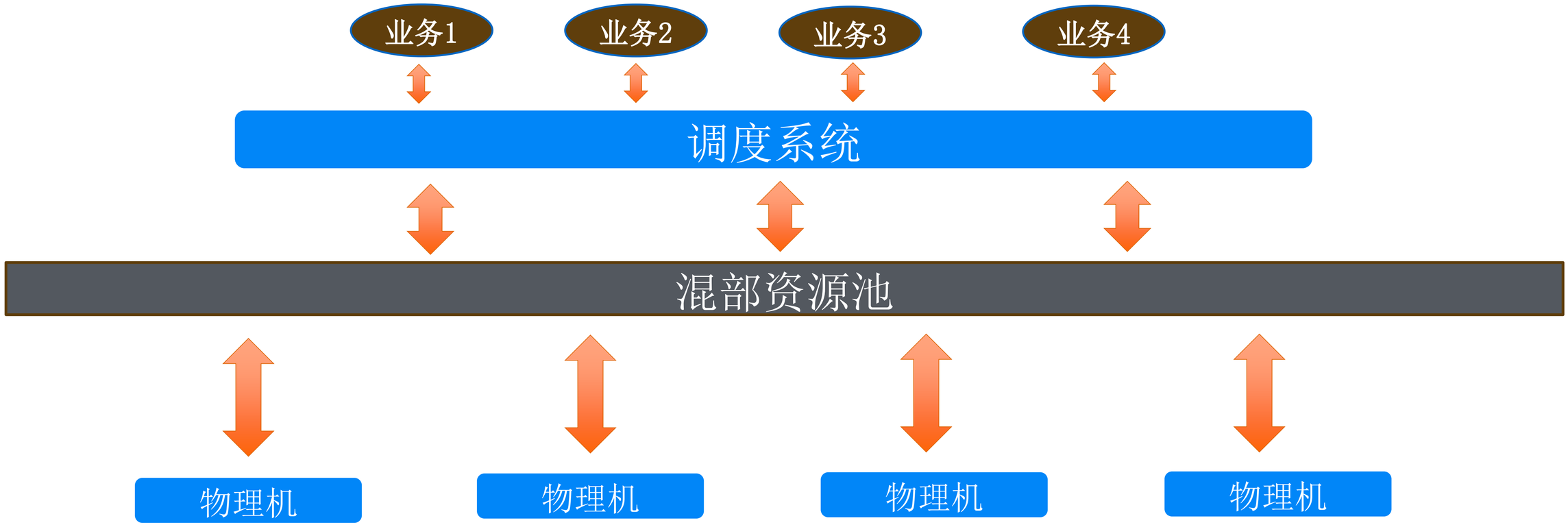
03 Memory cgroup priority

04 Per cgroup background reclaim

01 混部

混部

混部：将不同类型的业务调度到相同的物理资源上，通过硬件调配，资源隔离，灵活调度等控制手段，保障服务SLA，实现物理资源利用率的有效提升



02 超卖

超卖

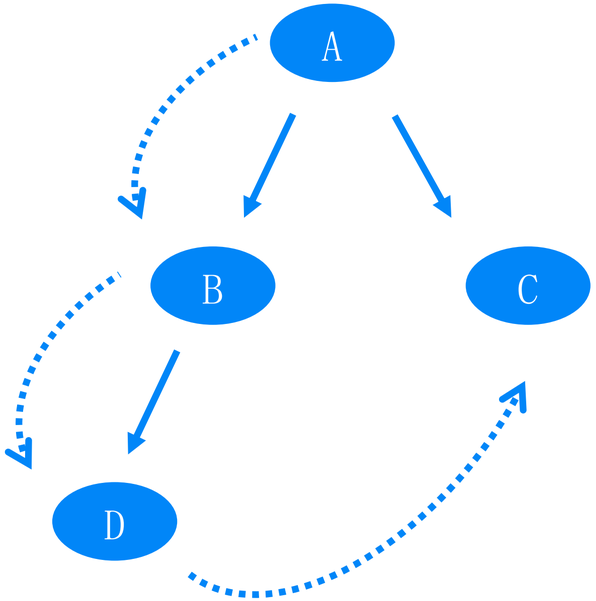
- 业务申请资源的总和大于物理机所拥有的资源
- 资源争抢
- 如何降低资源争抢对重要业务的影响
  - 内存
    - Memory cgroup priority
    - Per cgroup background reclaim

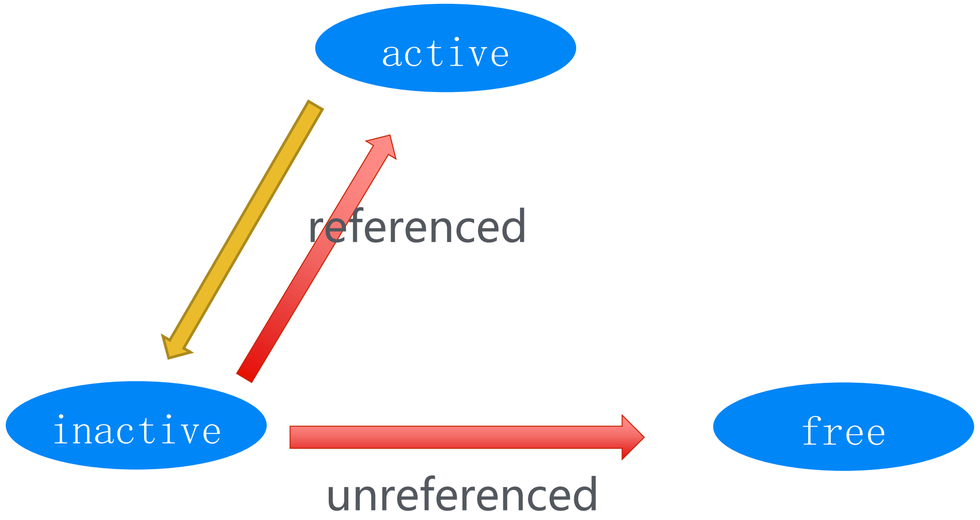
03 Memory cgroup priority

Memory cgroup priority

- 13个优先级：0~12，数字越高，优先级越高
- 作用于
  - memory reclaim(global reclaim & memcg reclaim)
  - out of memory(global OOM & memcg OOM)

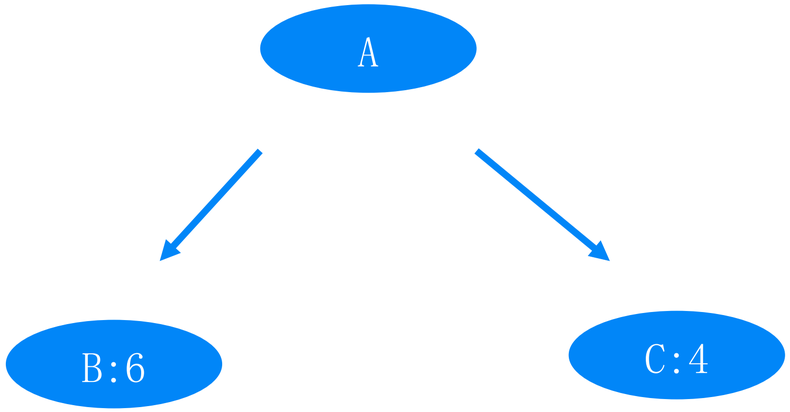
Memory reclaim





- 按照pre-order遍历扫描回收cgroup树
- 高优先级的cgroup一般情况下拥有较低的aging speed，从而其page不容易被回收，但在回收内存受阻的情况下，会提高其aging speed，以满足系统对内存的需求

Memory reclaim



Priority:

B: 6
C: 4

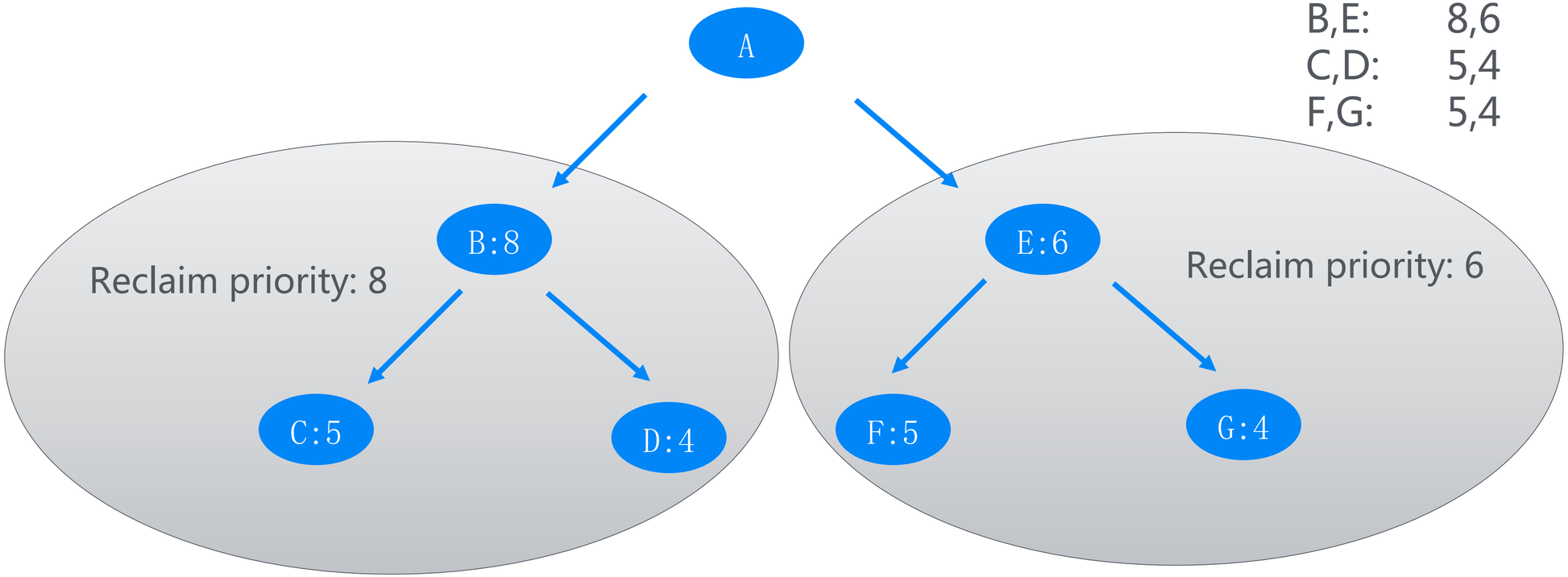
A usage > limit

B reclaim priority 6
C reclaim priority 4

Memory reclaim

Priority:

B,E: 8,6
C,D: 5,4
F,G: 5,4



A usage > limit

B,C,D reclaim priority == B priority:8
E,F,G reclaim priority == E priority:6

OOM

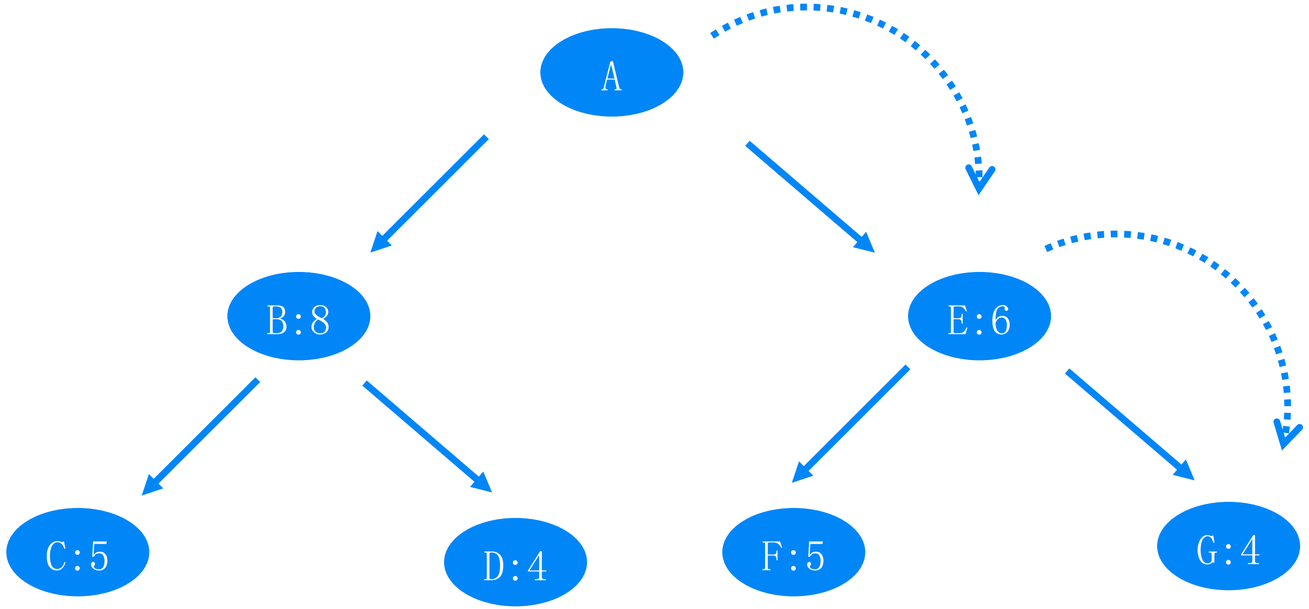
当发生OOM时，会按照优先级从高到底，从低优先级中选择受害者

Priority:

B,E: 8,6
C,D: 5,4
F,G: 5,4

A trigger OOM
B:8 > E:6 select E
F:5 > G:4 select G

same priority:
user defined strategy
chose max usage(default)



OOM

- 整组杀

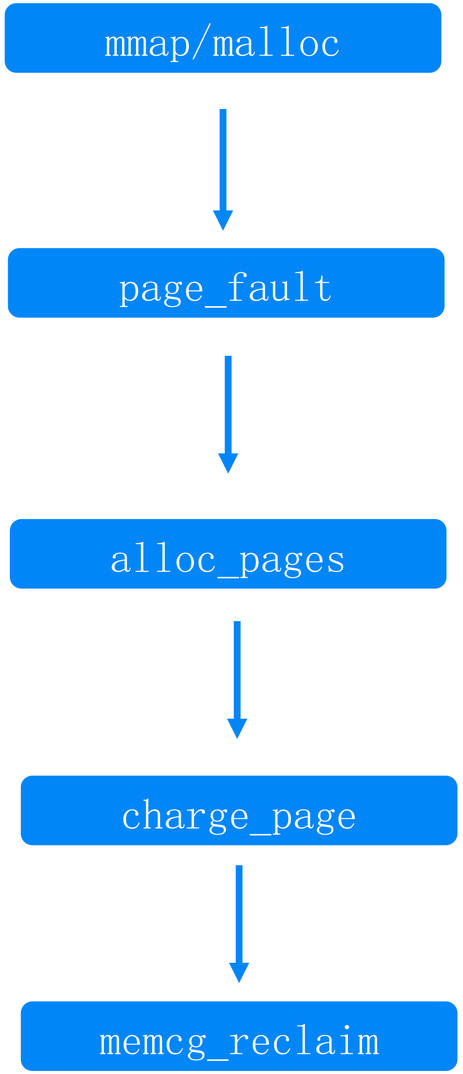
在一些应用场景下，当容器中某个进程被杀后，整个容器就无法正常工作，留下剩余的进程也没有意义。

因此我们提供整组杀的功能，当cgroup中某个进程被杀后，杀掉剩余其他进程。

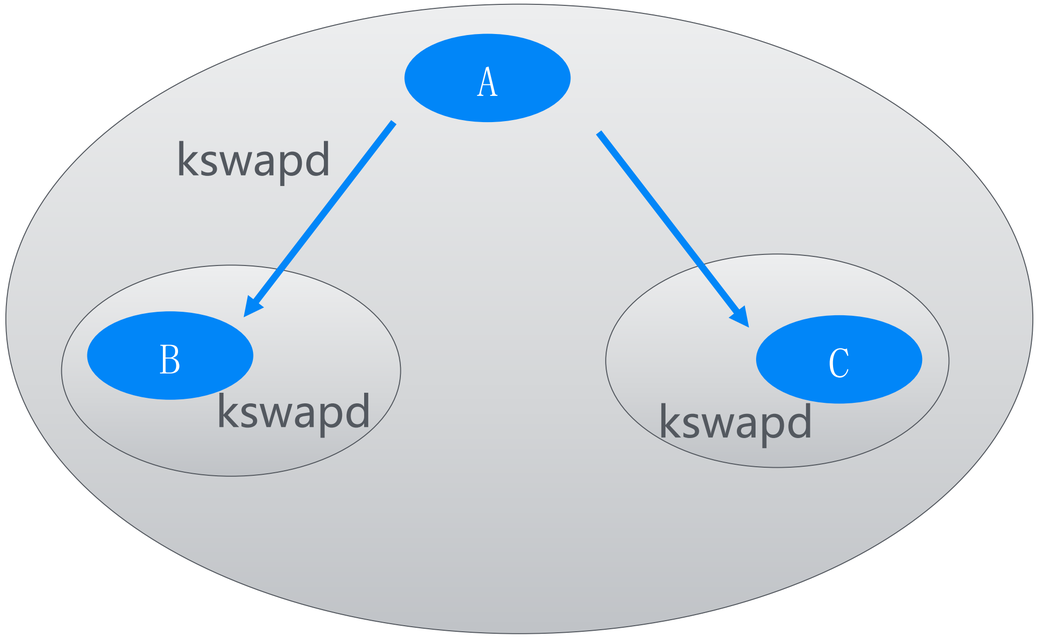
04 Per cgroup background reclaim

Per cgroup background reclaim

通过 per cgroup background reclaim 我们可以减少进入 memcg direct reclaim 的次数，从而减少 memcg charge的时间



Per cgroup background reclaim



- 每个cgroup都可配置一个相应的kswapd线程
- 当触发usage > low watermark时唤醒相应的kswapd线程
- 当usage < high watermark停止kswapd的回收，low, high watermark用户可配

Per cgroup background reclaim

超卖场景下：

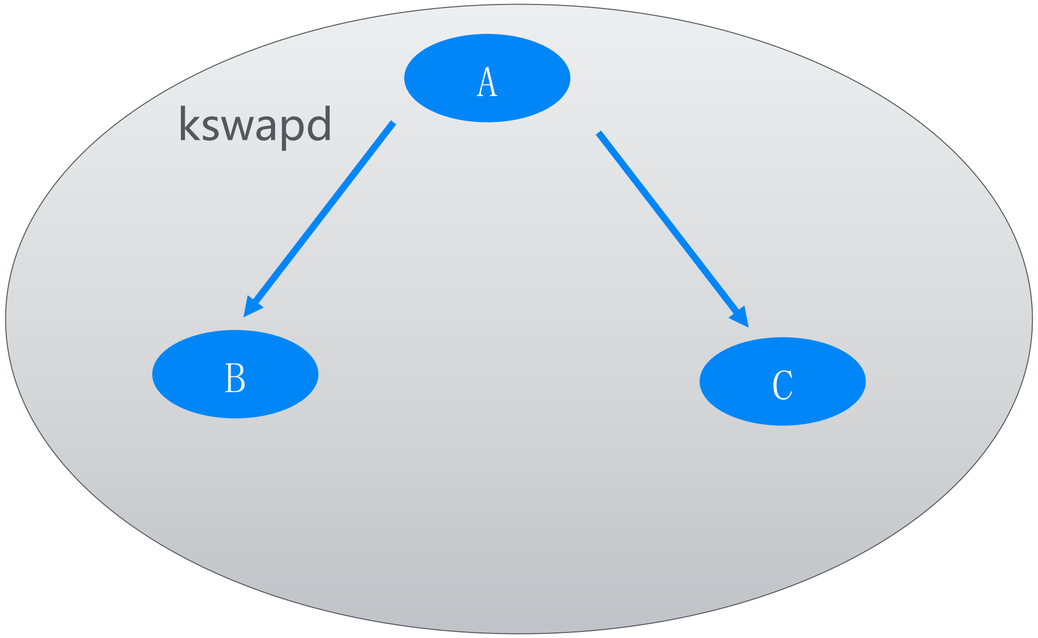
B组：低优先级
C组：高优先级
limit(B+C) > limit(A)

当usage(B+C) > limit(A) && usage(C) < limit(C)时,C在charge的时候会在A层触发direct reclaim，这时候B就对C产生了影响。

为了降低B对C的影响，我们可以enable A的kswapd线程：

1. A的kswapd线程进行background reclaim，可以减少C在charge page的时候进入 memcg direct reclaim的次数

2. 由于C的优先级高于B可降低其在background reclaim中受到的影响，使得reclaim的压力更多的放在B上



Reference

Per cgroup background reclaim

https://lwn.net/Articles/438246/

Alibaba Group 阿里巴巴集团

intel IBM IT大咖说 知识分享平台

Thanks

系统软件事业部　打造具备全球竞争力、效率最优的系统软件