Alibaba Group
阿里巴巴集团

阿里内部集群的内存超卖

阿里巴巴内核团队-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 backgroup reclaim

01 混部

混部

- 什么是混部

  将不同的业务跑在同一机器上

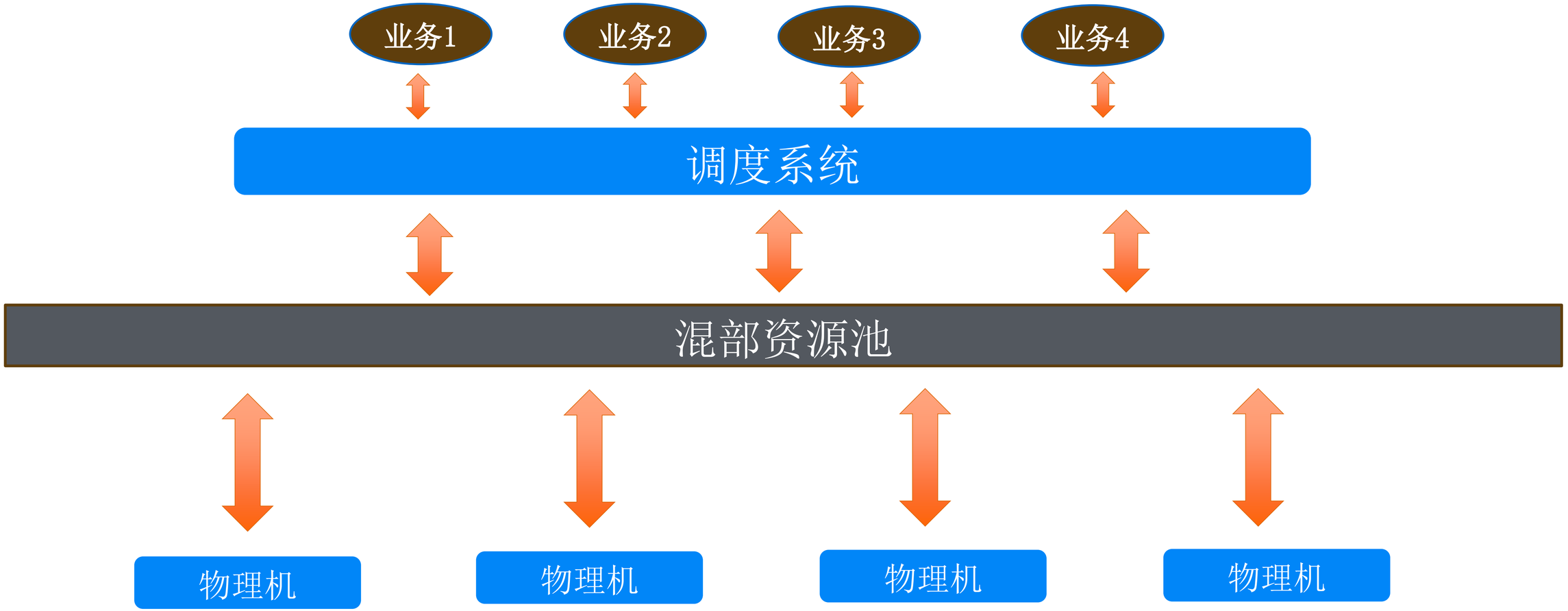
- 混部的目的

  提高物理资源的利用率，降低采购成本

- 混部的前提

  - 业务容器化
  - 调度系统

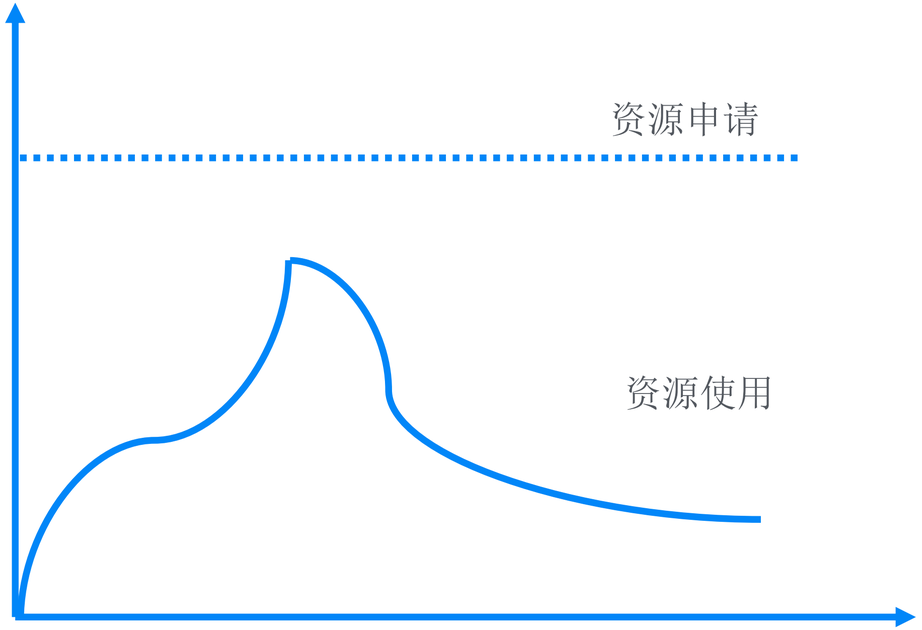
混部



02 超卖

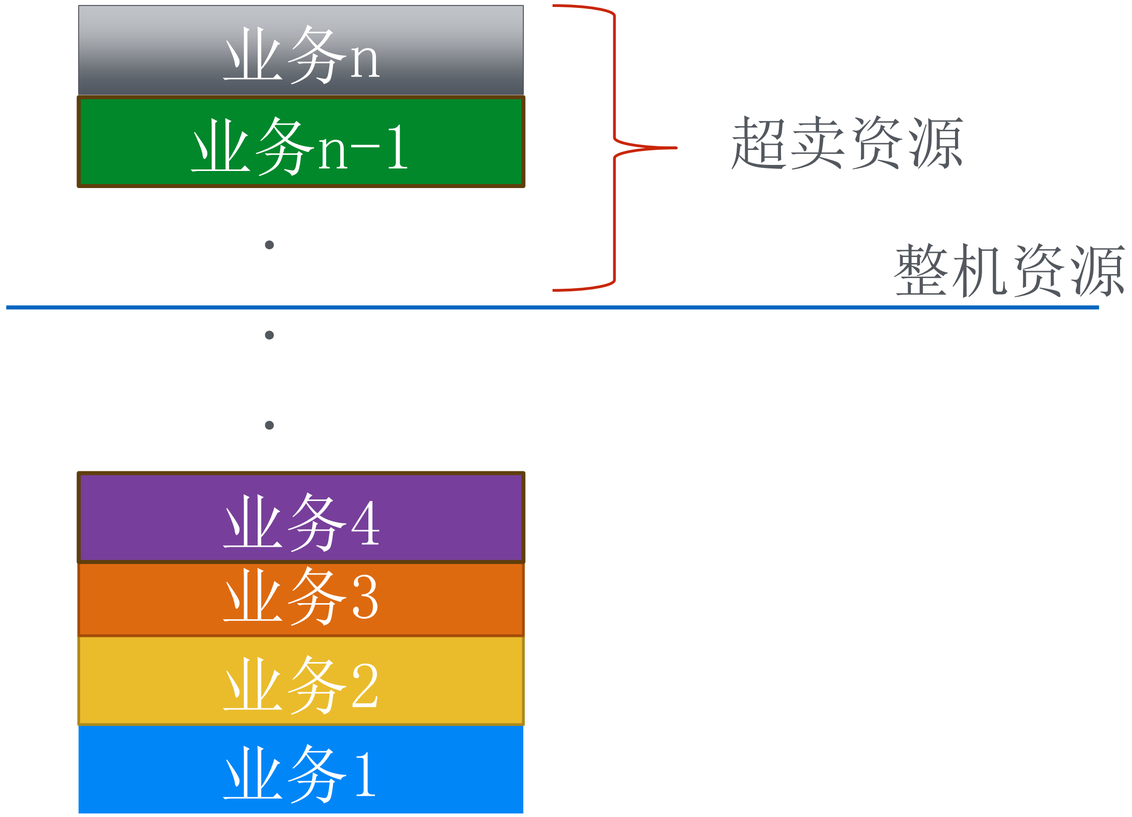
超卖

在对业务进行容器化过程中，业务由以前对物理机器的采购，转向对各项物理资源(cpu， memory， IO等)的申请



但在业务的实际运行过程中，我们发现业务对资源的使用，往往达不到它的申请，这就造成了资源的浪费

超卖



- 什么是超卖

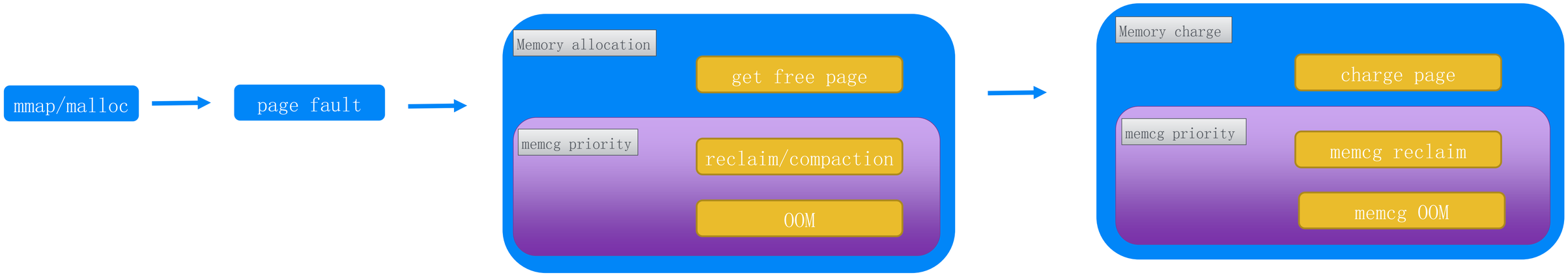
为了提高资源的利用率，我们引入超卖的概念，即在服务器上跑的业务所申请的资源的总和，超出了服务器本身所拥有的资源量

- 超卖面临的挑战

在超卖的过程中，会出现不同业务对资源的争抢的情况，这也是超卖的本质导致的：各个业务申请的资源超过了服务器实际提供的资源。所以如何保障重要业务的资源服务质量，成了我们的一个主要挑战

03 Memory cgroup priority

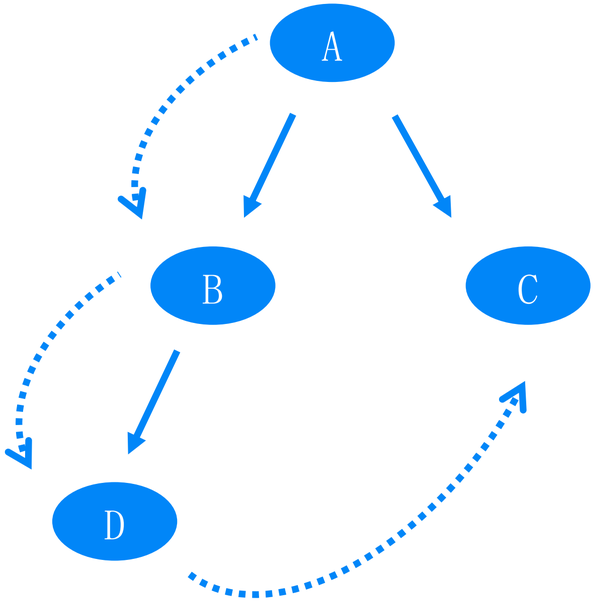
Memory cgroup priority

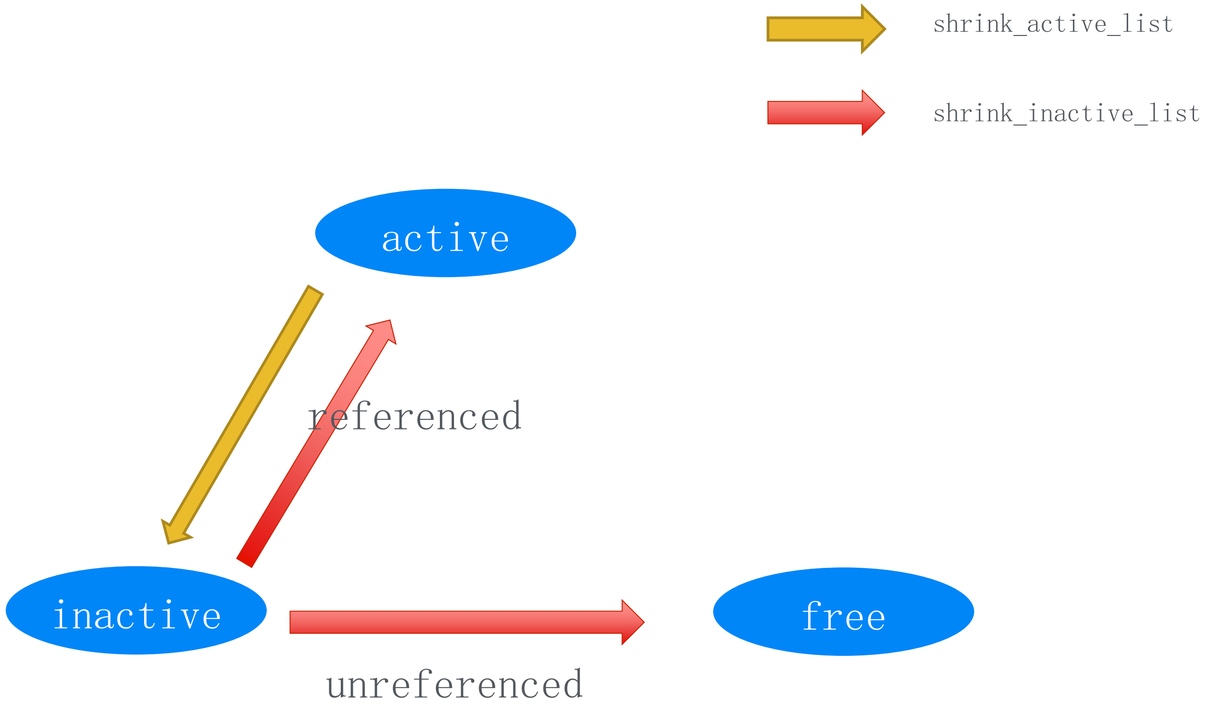


- 13个优先级：0~12，数字越高，优先级越高。高优先级的cgroup拥有更好的内存服务质量

- 作用于
  - memory reclaim(global reclaim & memcg reclaim)
  - out of memory (global OOM & memcg OOM)

Memory reclaim



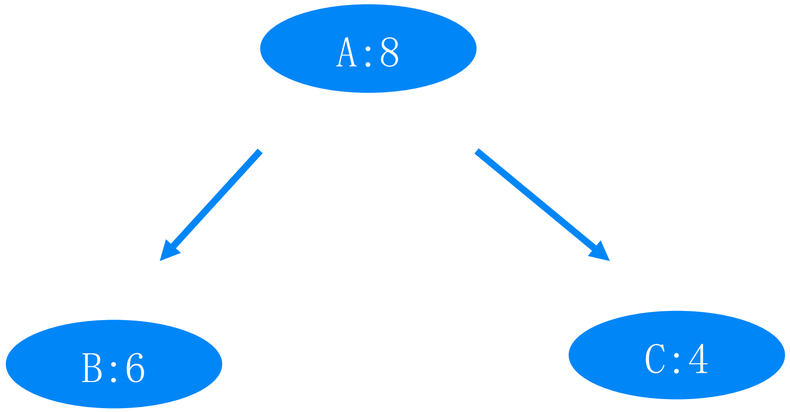


- 按照pre-order遍历扫描回收cgroup树
- 高优先级的cgroup一般情况下拥有较低的aging speed，从而其page不容易被回收，但在回收内存受阻的情况下，会提高其aging speed，以满足系统对内存的需求

Memory reclaim

Priority:

A: 8
B: 6
C: 4



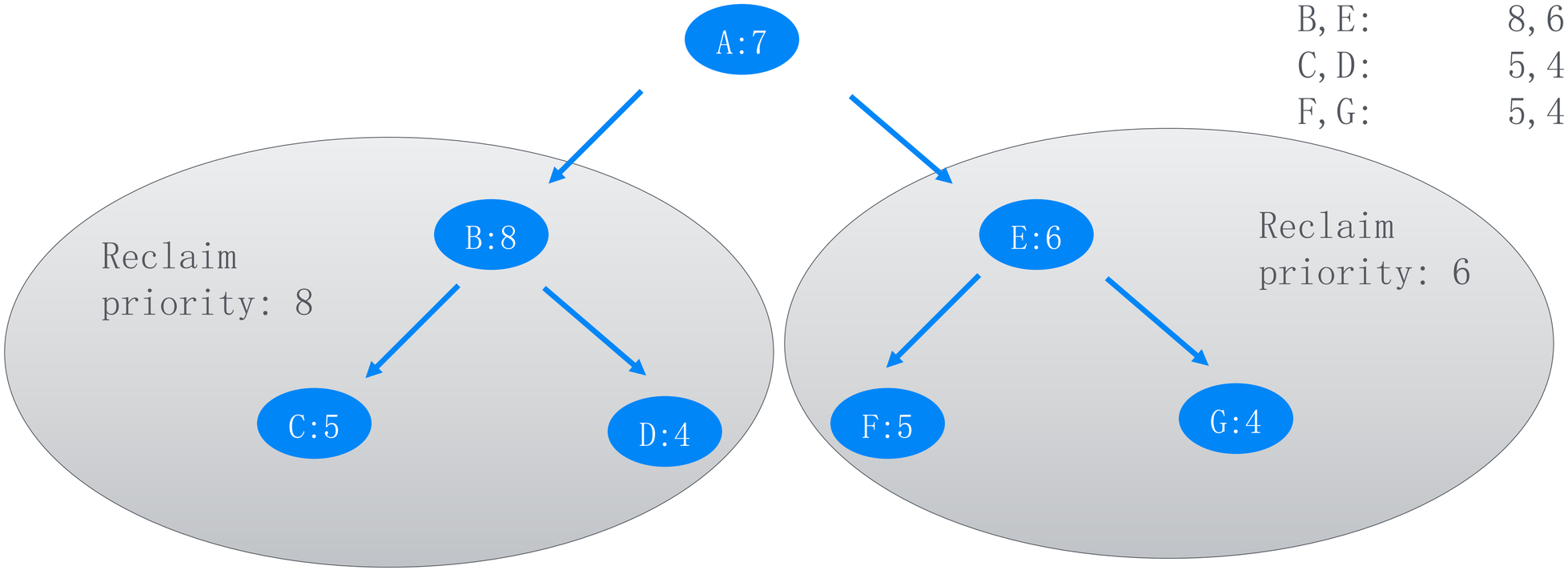
A usage > limit

A reclaim priority 0
B reclaim priority 6
C reclaim priority 4

Memory reclaim

Priority:

A:	7
B, E:	8, 6
C, D:	5, 4
F, G:	5, 4



A usage > limit

A reclaim priority: 0

B, C, D reclaim priority == B priority:8

E, F, G reclaim priority == E priority:6

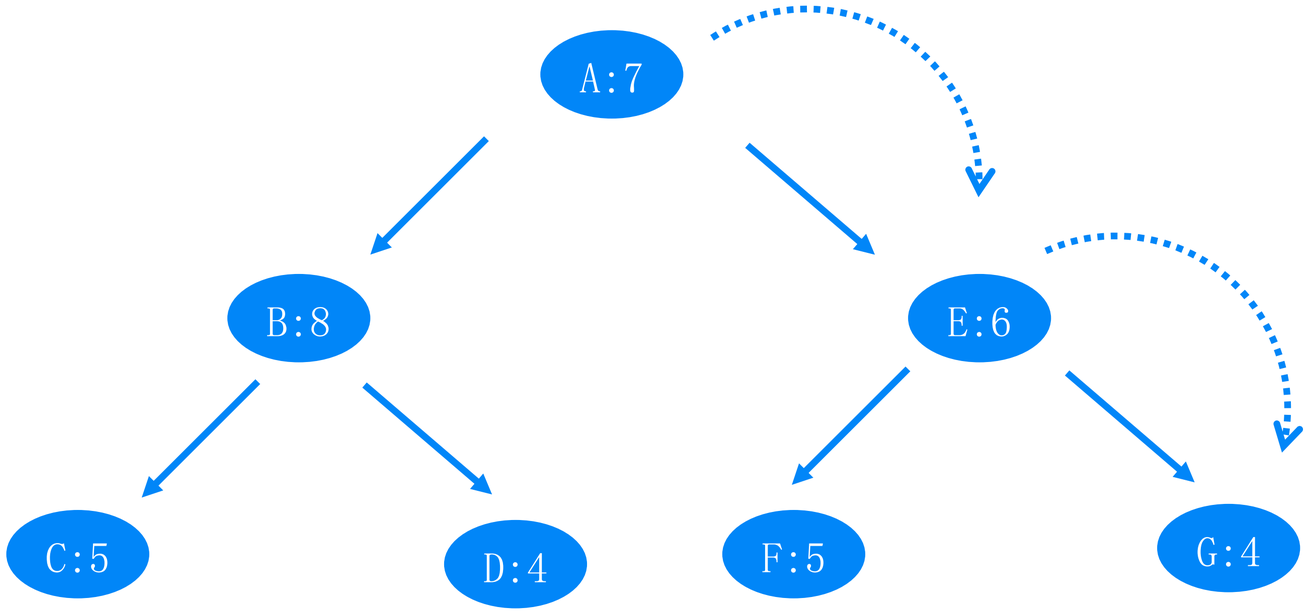
OOM

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:	
A:	7
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。

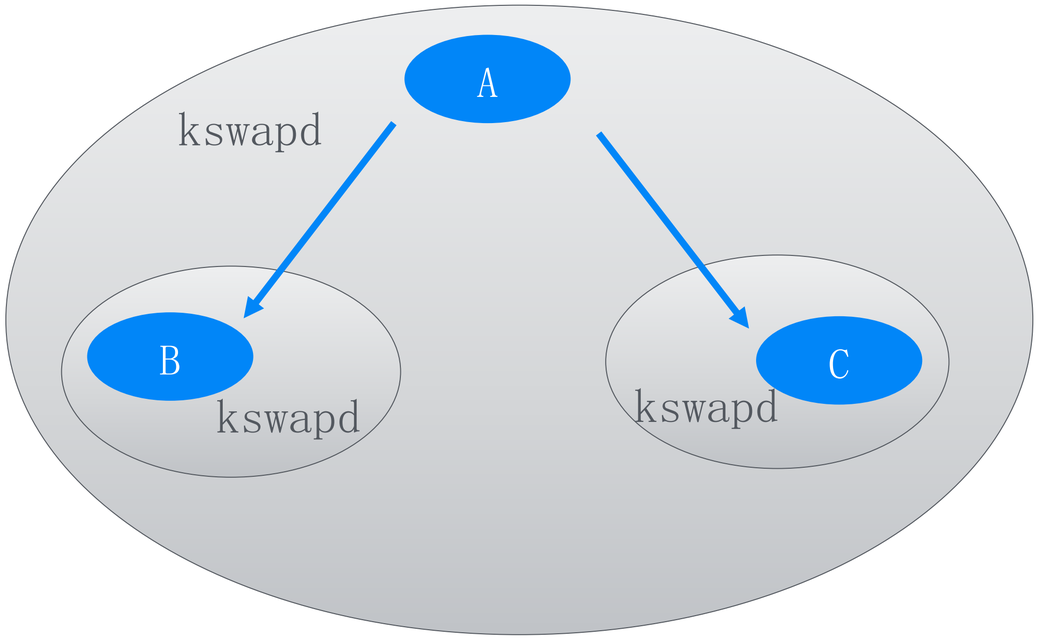
04 Per cgroup background reclaim

Per cgroup background reclaim

原作者: Ying Han

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

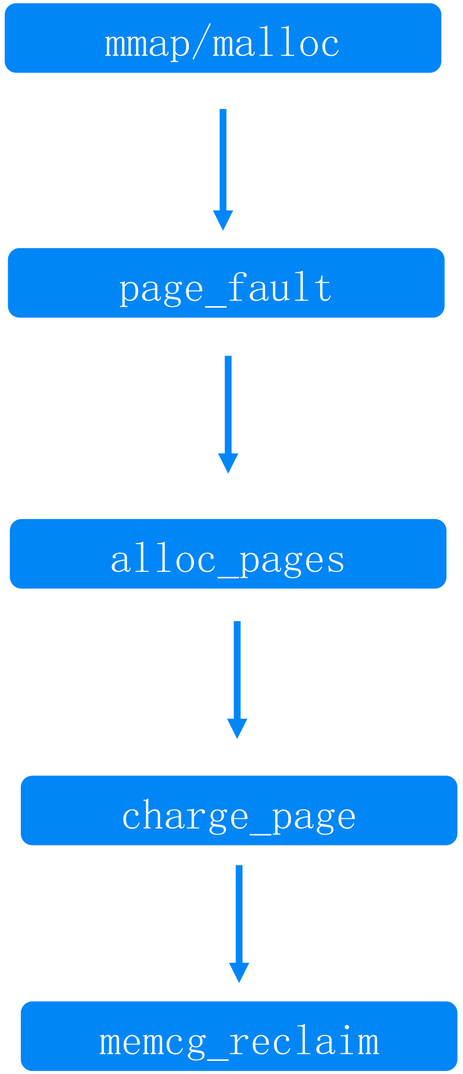
Per cgroup background reclaim



- 每个cgroup都可配置一个相应的kswapd线程
- 当进入到memcg direct reclaim时唤醒相应的kswapd线程
- 当usage降到high watermark时停止kswapd的回收，high watermark用户可配

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

超卖场景下：

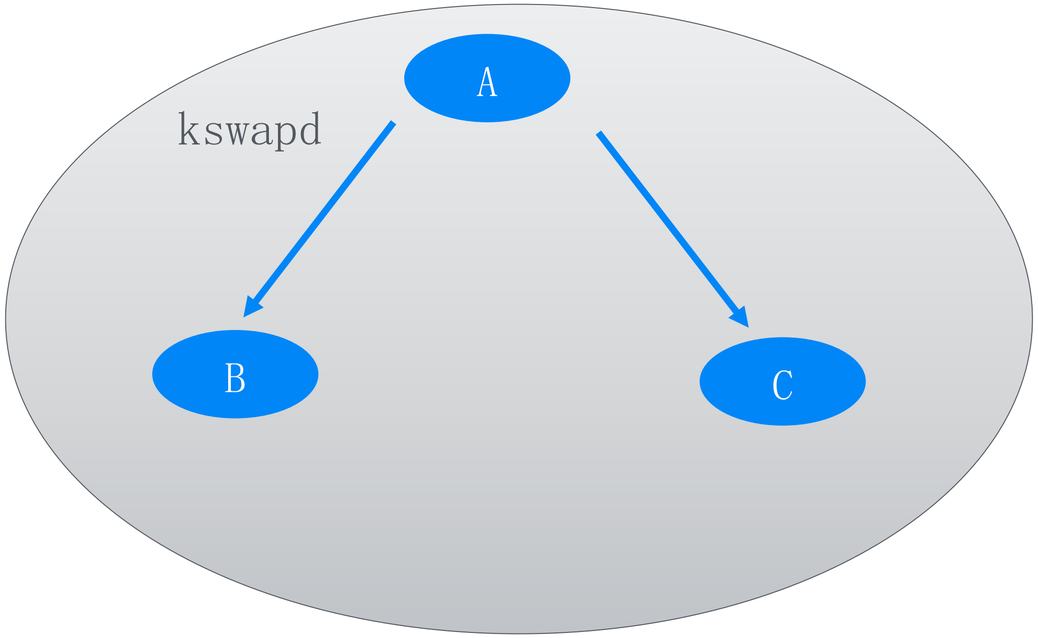
B是超卖组　　低优先级
C是正常组　　高优先级
limit(B+C) > limit(A)

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，这时候超卖组就对正常组产生了影响。

为了降低超卖组对正常组的影响，我们可以enable A的kswapd线程：

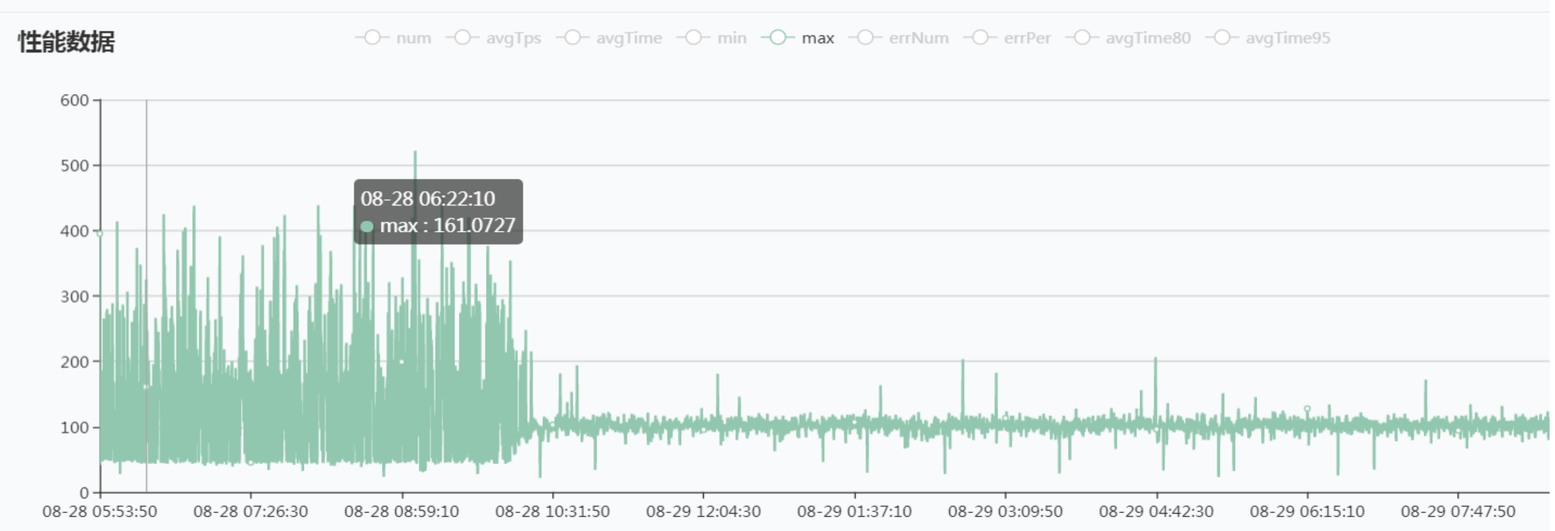
1. A的kswapd线程进行background reclaim，可以减少C在charge page的时候进入 memcg direct reclaim的次数

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



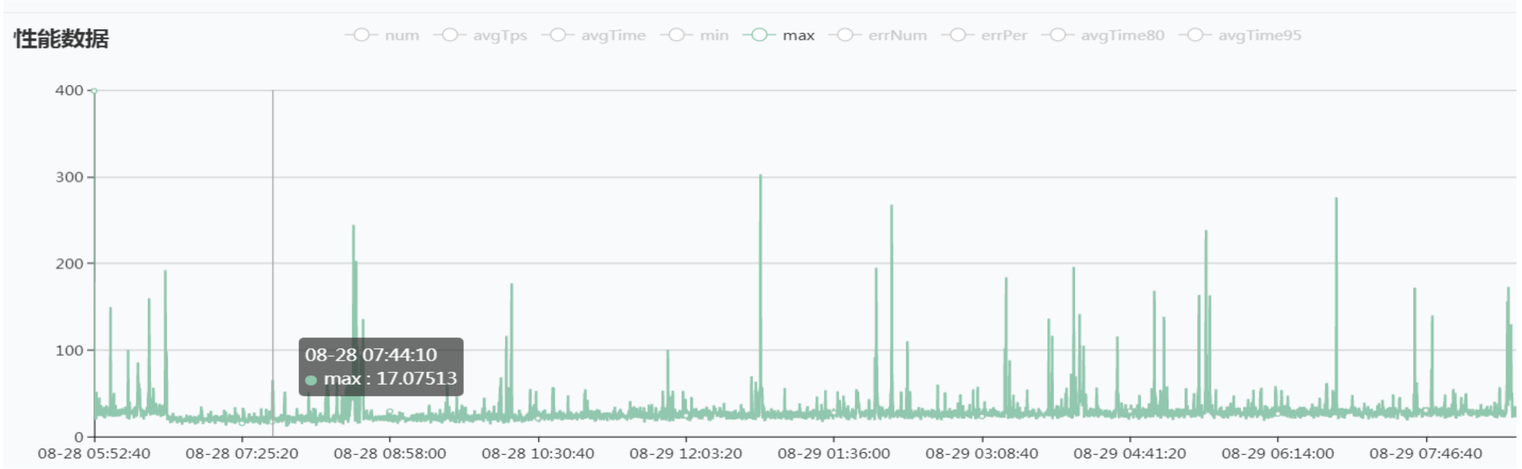
Per cgroup background reclaim

未开启 per cgroup background reclaim 最大RT



开启per cgroup background reclaim，每10s的所有请求的最大响应时间平均值在20ms左右，未开启的平均值在100ms左右

开启 per cgroup background reclaim 最大RT



Alibaba Group
阿里巴巴集团

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