腾讯tlinux调度算法改进

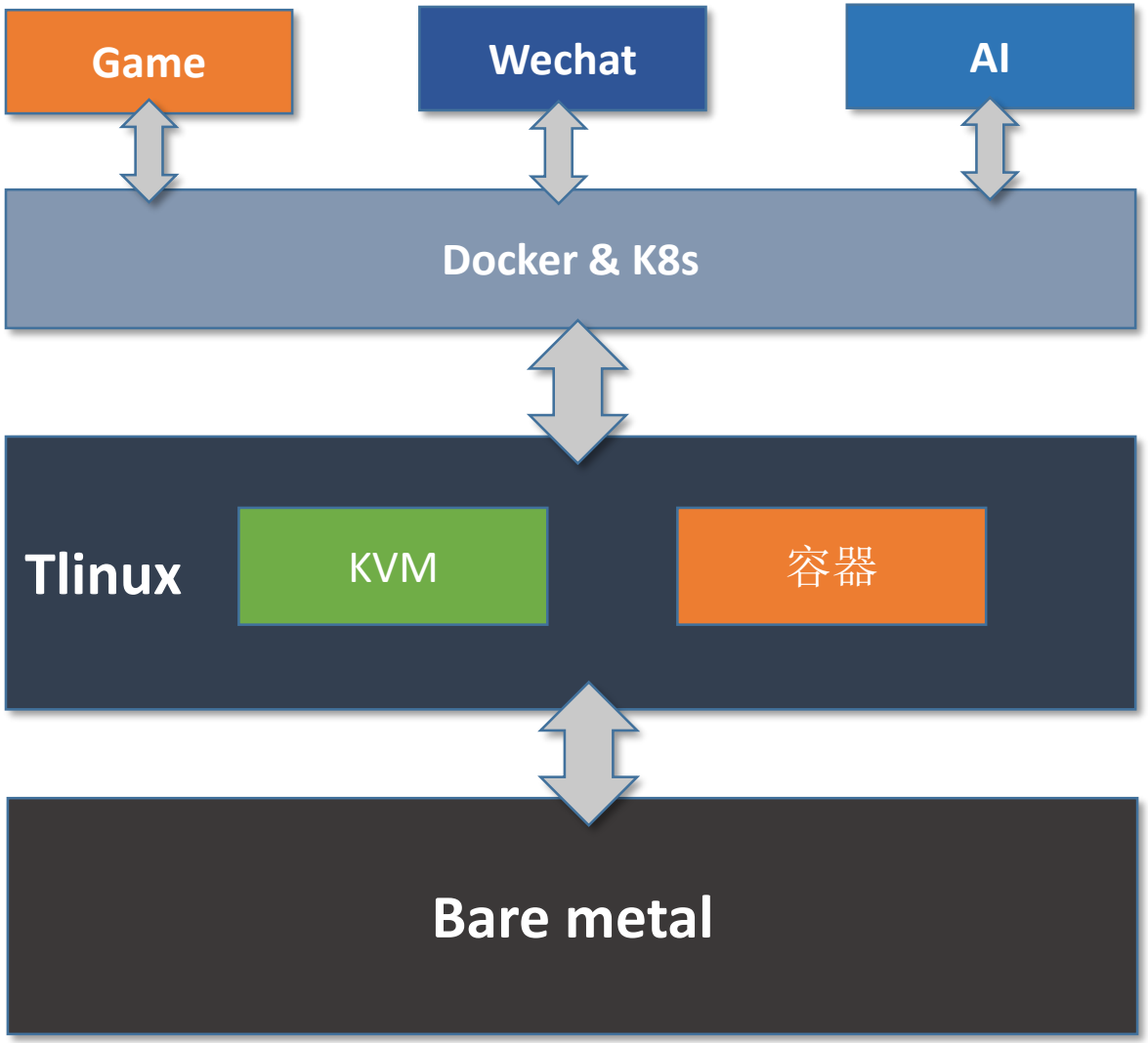
演讲者：高小明(newtongao@tencent.com)

Tencent

- 背景
  - 随着公司业务的发展，服务器数量以及单机规格都在增大,服务器的空闲率越来越高，WXG希望通过离线&在线混部的方式来提高服务器利用率。
  - 业界提高服务器利用率的方式，有虚拟化和容器, IDC内部更适合用容器。
- 现状
  - Linux内核默认的cgroup方式只能用于延迟不敏感业务。
  - Cgroup方式在离线进程量多的时候，自身就会带来很大cpu开销。
- 挑战
  - 需要一种方案，既对在线造成很小的影响，又能部署离线计算，提高整体cpu使用率。
  - 目前业界还没有这种方案。

背景（补充图片）

腾讯内部主要使用kvm+容器，或者在裸金属上直接使用容器。

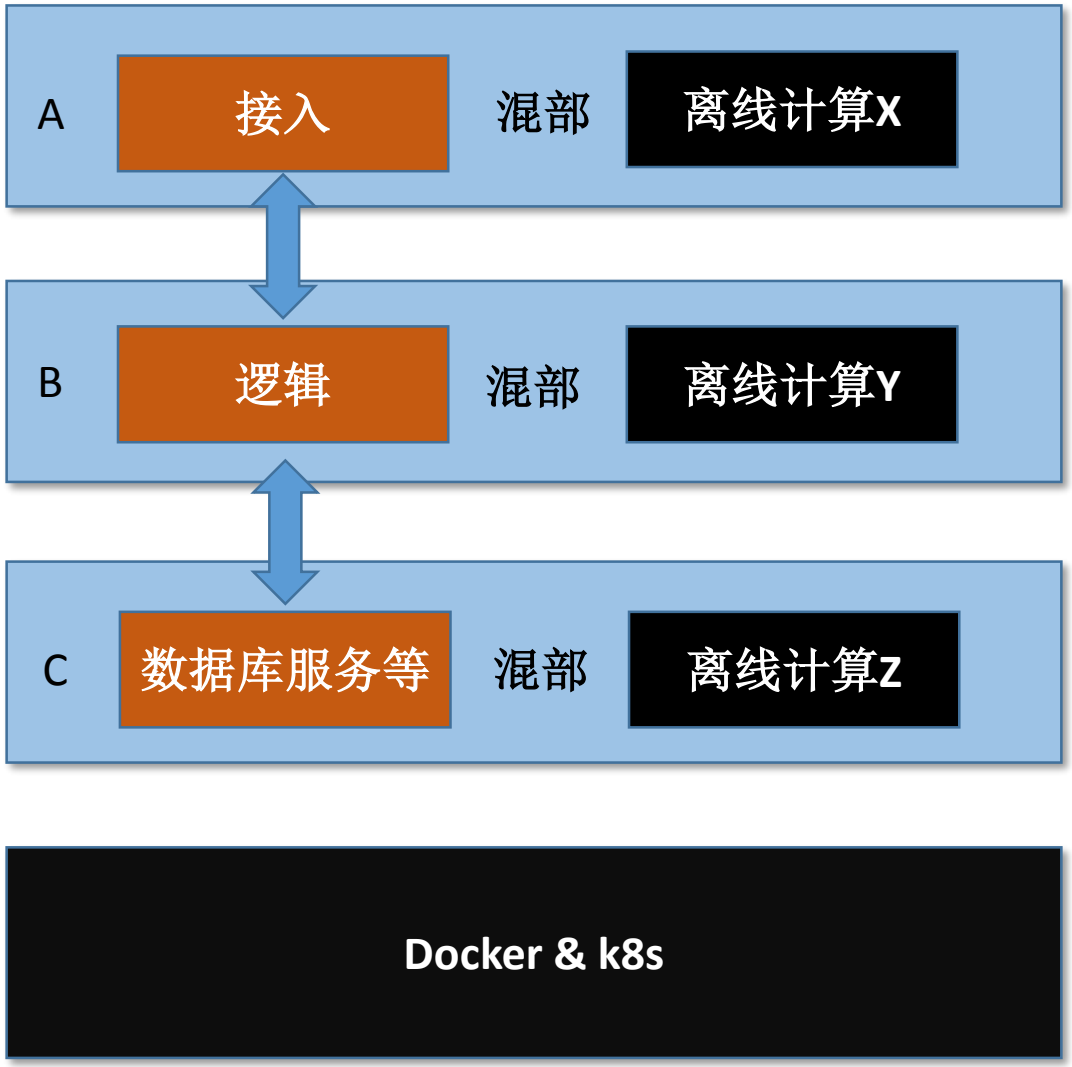


背景（补充图片）

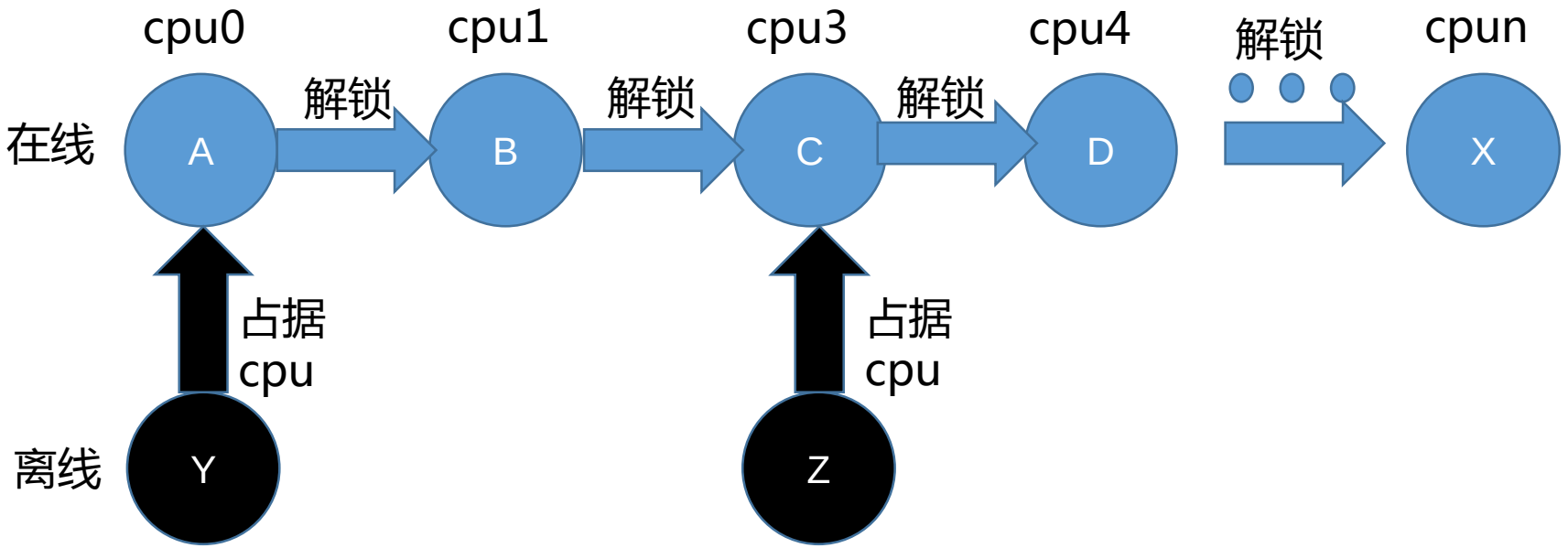
在线服务
- 高延迟敏感
- 高可靠

离线业务
- 低延迟敏感
- 大多数是计算型

混部的目标&意义
- 提升服务器cpu利用率。
- 节约计算成本。
- 计算资源更加弹性。



经典问题

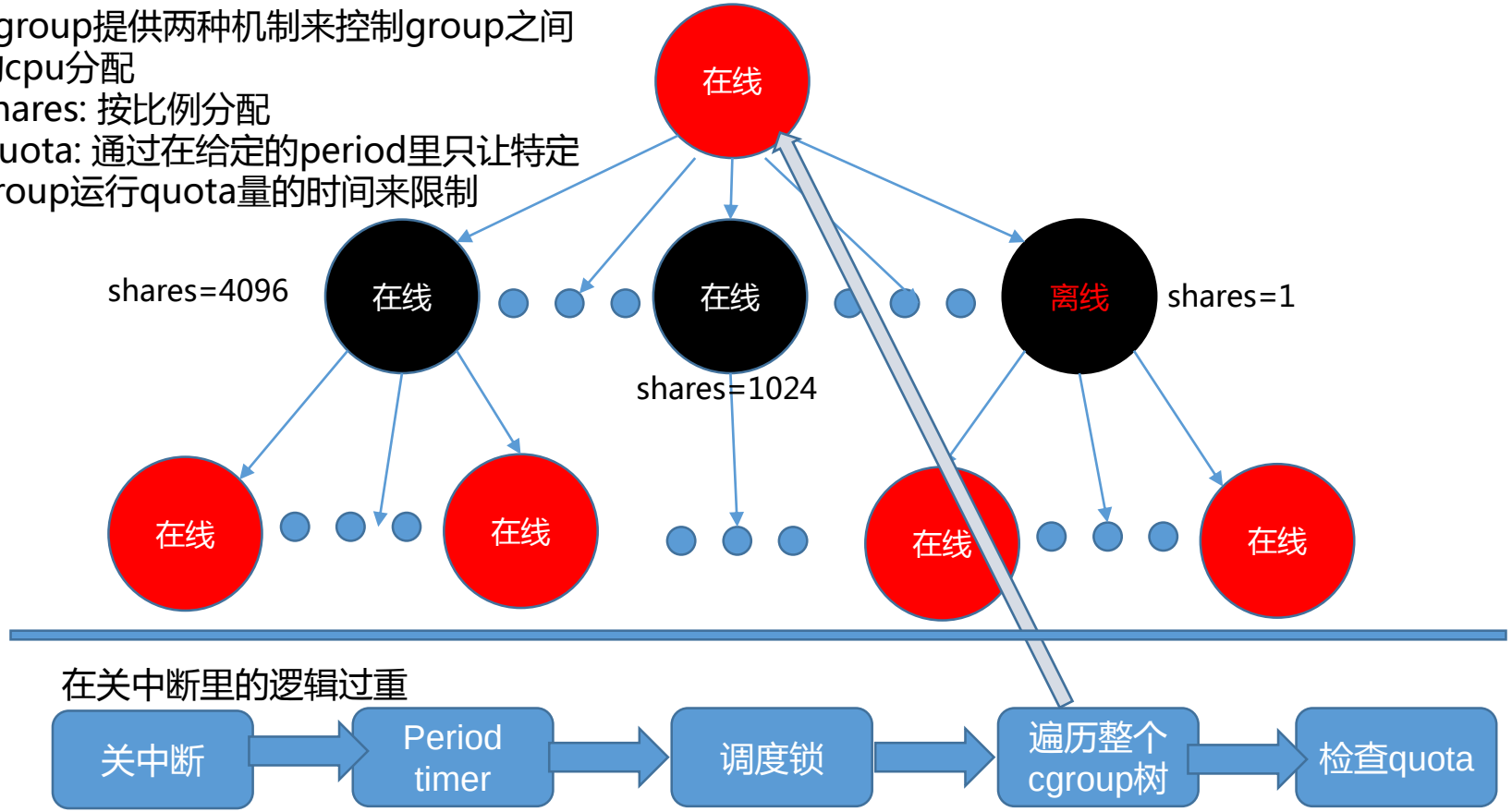


延迟传播效应

- 在循环依赖（很常见）的情况下，一个点的延迟会传播到下游结点。
- 会导致cpu空转，资源浪费。
- 只有保证在线进程的每个结点都不会被离线进程影响才能解决。
- 这种现象在虚拟化的spinlock里也常见。

Cgroup & Cfs 现有问题-Cfs自身问题

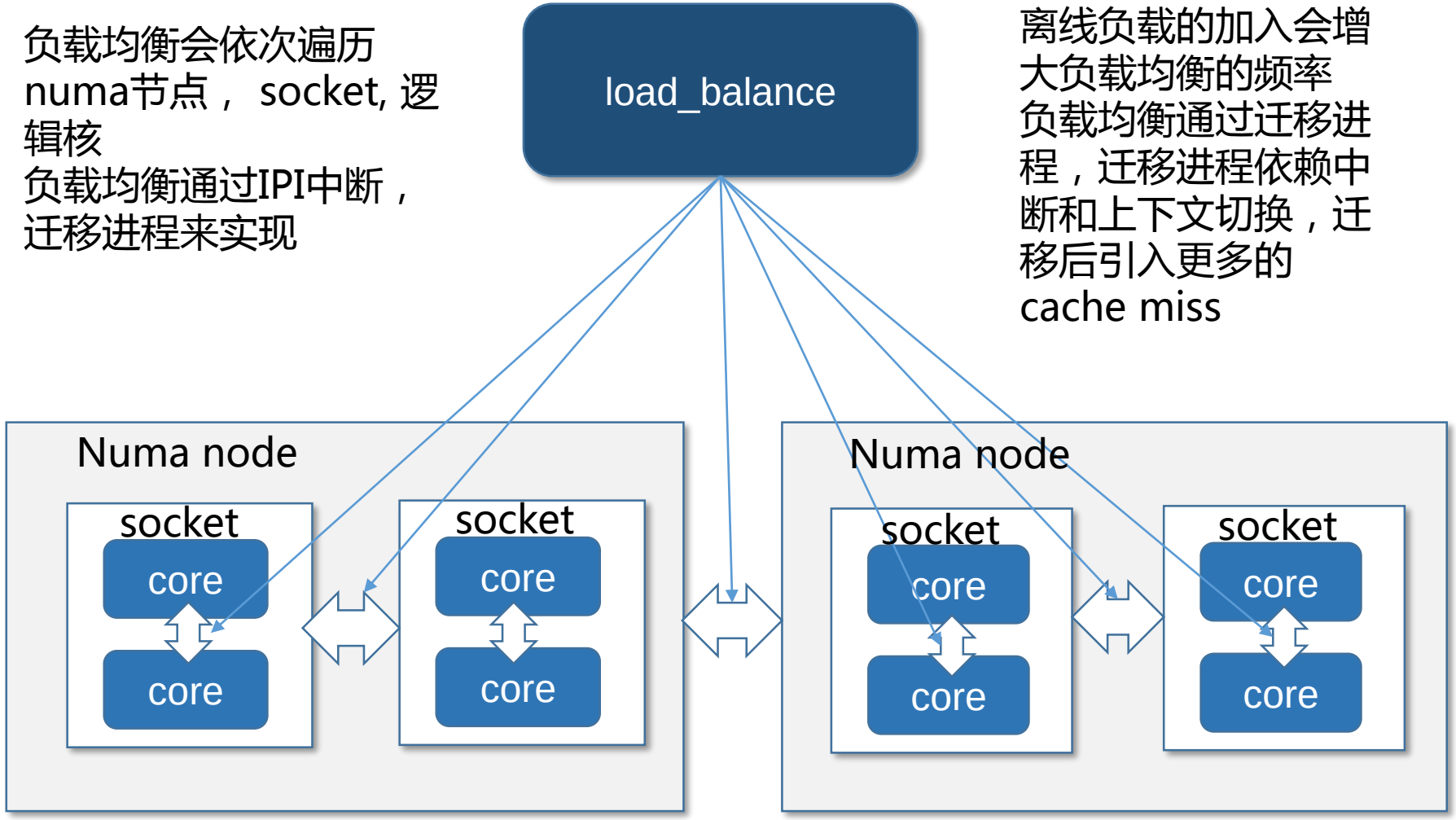
Cgroup提供两种机制来控制group之间的cpu分配
Shares: 按比例分配
Quota: 通过在给定的period里只让特定group运行quota量的时间来限制



Shares和period/quota机制都存在同样的问题：在线进程不能及时抢占离线进程的CPU，但其实这也是CFS算法决定的，CFS的初衷就是为了公平不是为了抢占。
period/quota会比shares好一些，可以通过将period设的比较小来更细的限制离线进程的对在线进程的影响，但是比较小的peroid又会带来更频繁的时钟中断，并且在关中断里加锁遍历整棵树。
现网就出现过多起由于period设置过小导致softlockup, 整机几乎不可用的案例。

Cgroup & Cfs 现有问题--负载均衡问题

负载均衡会依次遍历numa节点， socket, 逻辑核
负载均衡通过IPI中断，迁移进程来实现



离线负载的加入会增大负载均衡的频率
负载均衡通过迁移进程，迁移进程依赖中断和上下文切换，迁移后引入更多的cache miss

每个core还包含超线程， 图就不太细化了。

Tlinux的方案

◆ 面临问题

-敏感型在线业务无法混部
-Cpu利用率低

◆ 方案目标

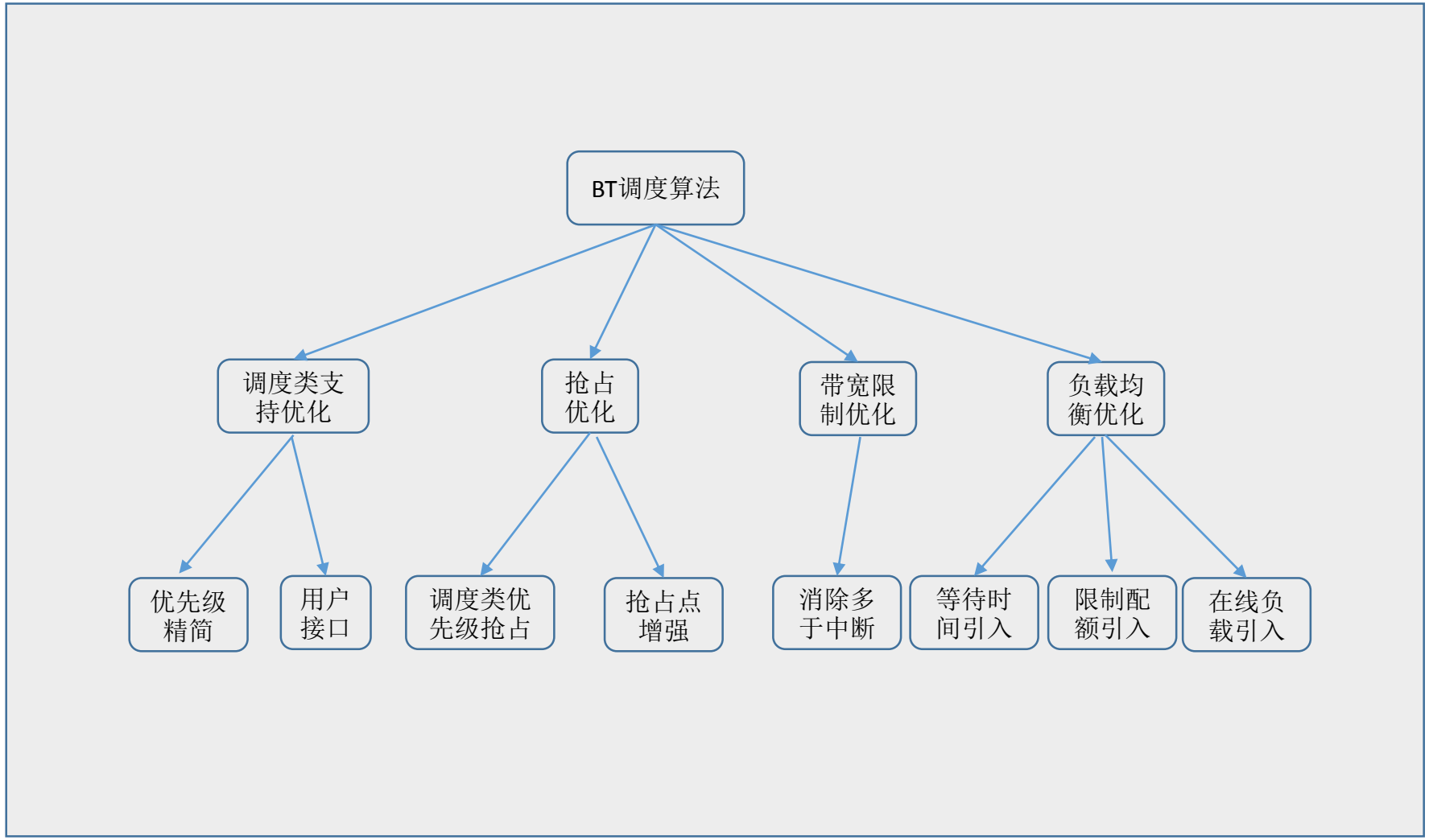
-不影响在线业务
-提升整机cpu利用率

◆ 方案

-结合离线业务的特点，提出一种新的度类，优先级低于cfs，从而做到离在线能够混部，不影响在线业务的目的，并优化均衡方案，减少在线业务对离线业务的影响，达到提升整机cpu利用率的目的。

BT调度算法框架

- 调度类优化
- 抢占优化
- 带宽限制优化
- 负载均衡优化



◆ 带宽限制目标

-限制离线cpu占用率

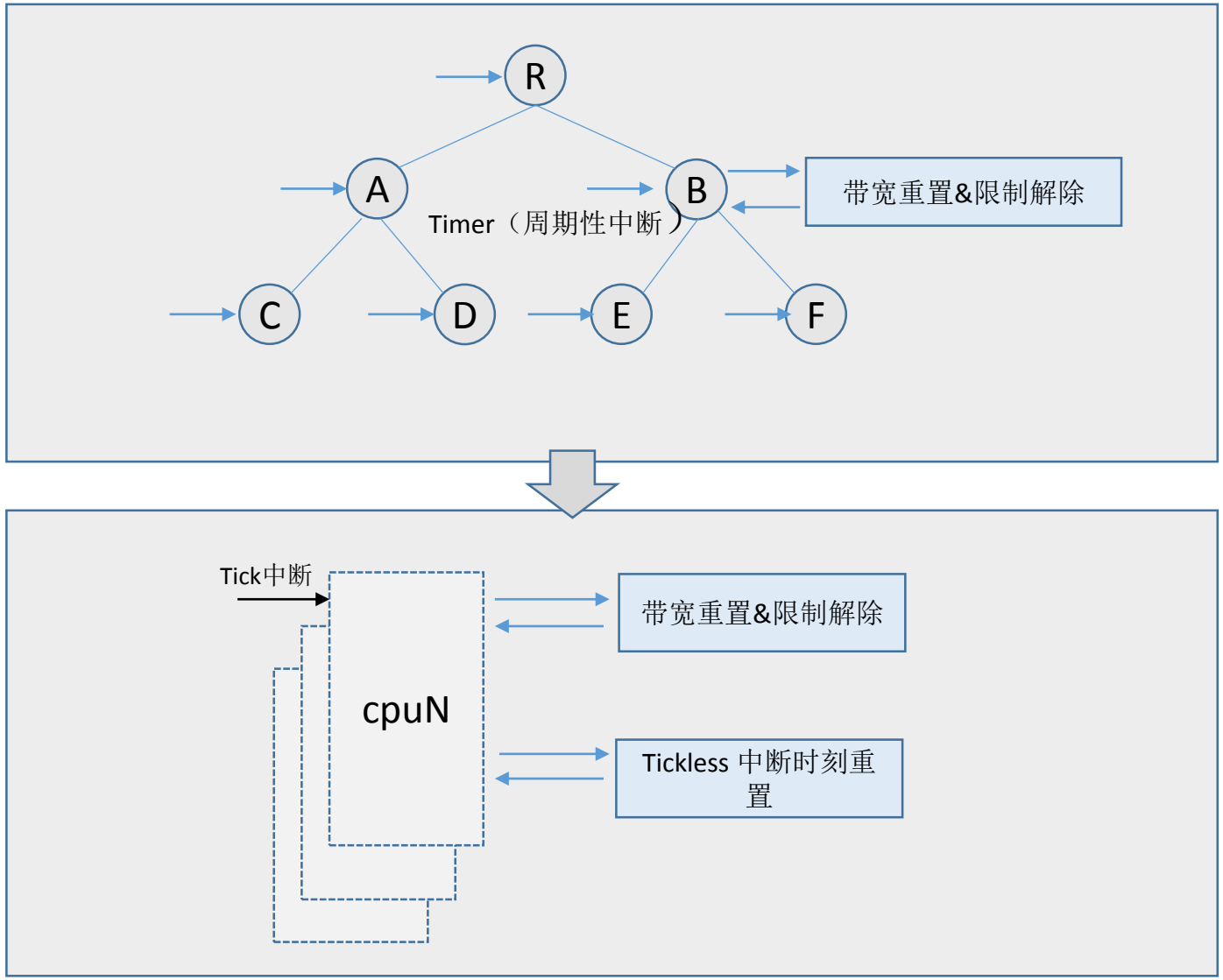
-不带来影响在线的开销

◆ Cfs带宽限制问题

-增加额外中断影响性能

◆ BT带宽限制方案

-复用tick中断，减少额外中断的影响



社区初稿提交

腾讯开源目录
https://github.com/Tencent

后续计划
后续我们会整体开源
包括docker和内核
并附上详细说明文档

Tencent