土豆视频CDN的应用与实践

吴岷@土豆网

Contents

- 土豆视频CDN
- 我们遇到的困难
  - 单台服务能力有限
  - 调度困难
- 我们的做法
  - 提高单台服务能力
  - 动态调度每个请求
- 实时调度系统
- 分享一些运维相关信息

土豆视频CDN简介

- 只提供视频服务（HTTP协议），包括点播和直播
- 和我们提供图片服务的CDN的异同点在于
  - 相同点：
    - 尽量离用户近
  - 不同点
    - 性能：（土豆）图片CDN是disk I/O bound的应用；而视频CDN一般是网络I/O bound应用
    - 同步：图片CDN采用pull模式同步文件，每个点都有所有文件；而视频CDN采用push模式，每个点都只有一部分文件
    - 调度：图片CDN依赖DNS，而视频CDN依赖动态调度

我们遇到的困难 – 单台服务能力

- 我们发现lighttpd并发超过600，性能就会快速下降
- 假定要满足300万同时在线，至少需要5千台服务器。

我们遇到的困难 – 调度困难

- 机房多：8Gb带宽的机房，假设每个连接是400kbps的下载速度，同时可以服务的用户数是8Gb/400kb = 20000个，需要一百多个机房才能满足几百万同时在线
- 文件不全：土豆网现在有4000多万个视频，假设视频的平均大小是25M，总共的存储空间是1Peta，即需要1000块1TB的硬盘

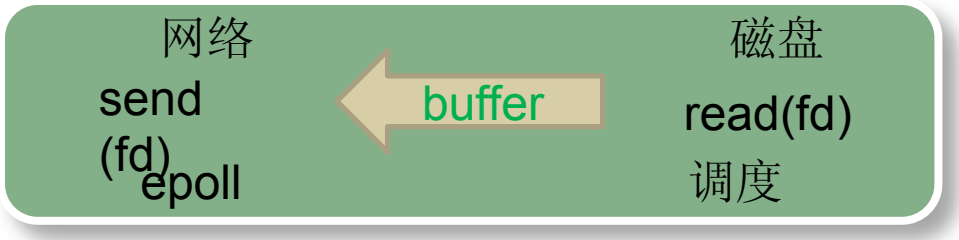
我们的做法

- 提高单台服务能力
- 开发动态调度系统来调度每个请求，解决机房之间负载平衡的问题

提高单台机器的服务能力

- 服务器配置：2core+10sata+4G内存+2个千兆网卡
- 分析600并发瓶颈的原因：lighttpd使用sendfile，disk I/O由操作系统调度
- 可能的改进点：
  - 能否控制一次磁盘读取大小?
  - 能否根据不同的硬盘做优化?
  - 能否根据网络层数据量作调度优化?

提高单台机器的服务能力



- 我们的做法：
  - 使用read/send来读磁盘和发送数据
  - 增加单次read读取大小
  - 自己实现对读磁盘操作的调度
    - 针对硬盘做调度优化
    - 针对网络层buffer中的不同链接数据量的差异做调度优化
- 效果：
  - 单台服务器达到2.5k个链接，跑满两张网卡

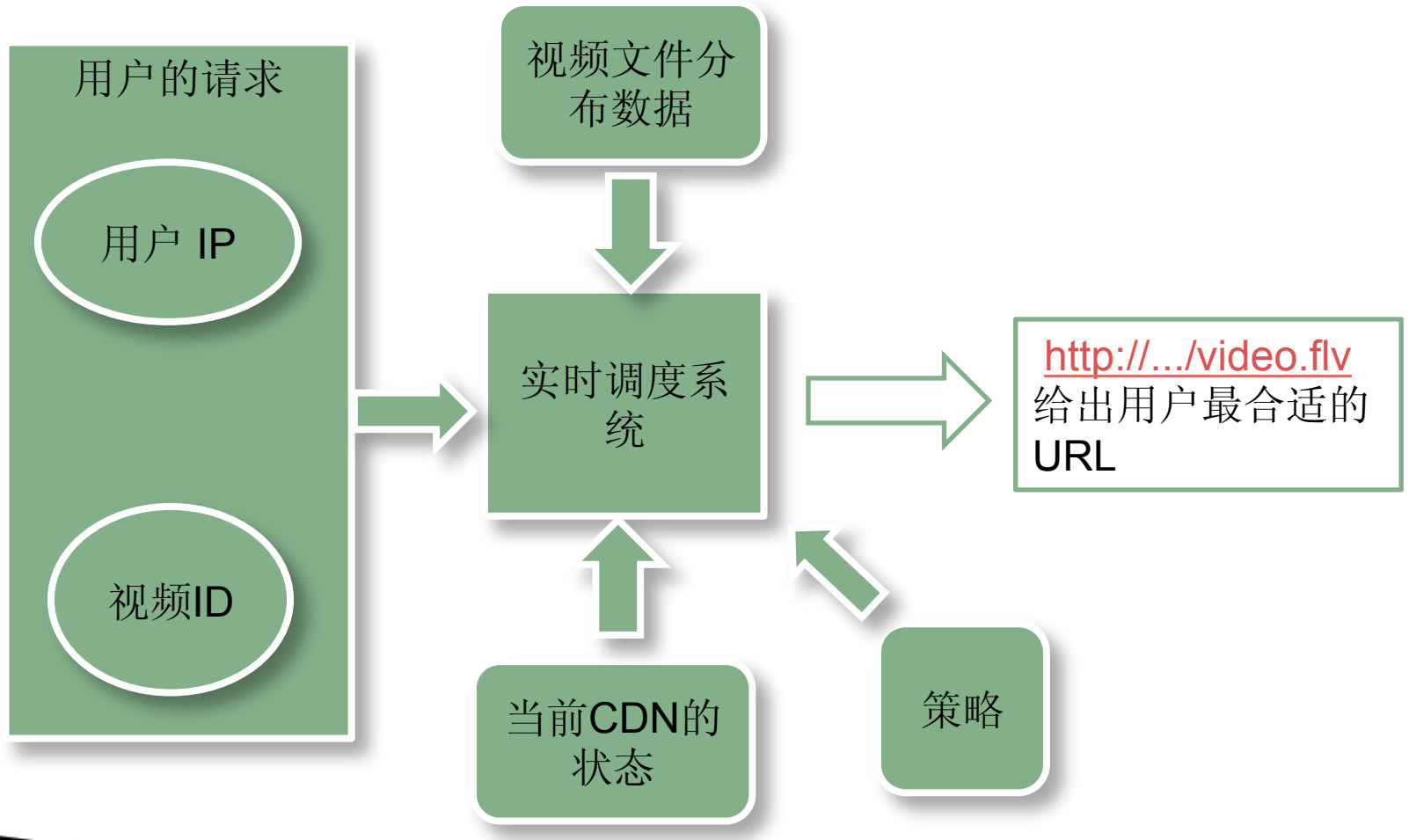
实时调度系统

1000公里

1公里

去哪儿吃饭呢？

实时调度系统 – 数据



实时调度系统－ 挑战

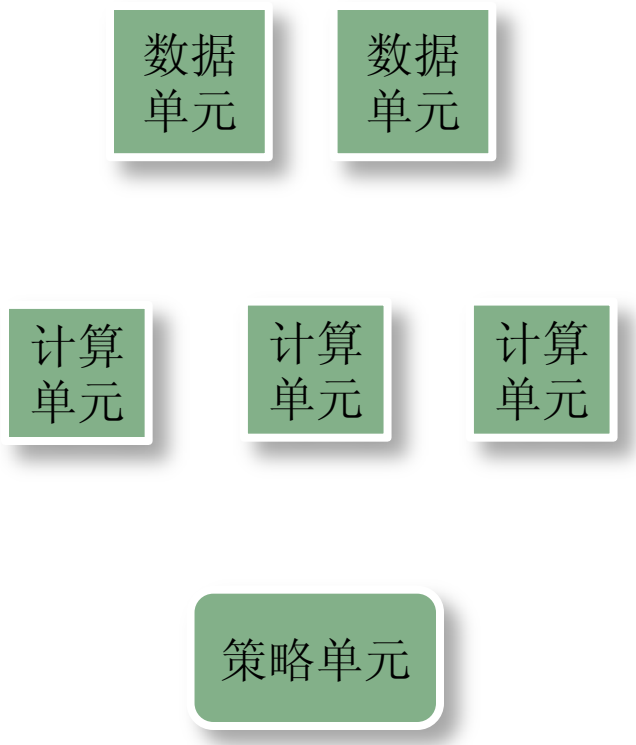
- 全国上百个节点，几百G带宽，每个节点出口带宽不同，文件也不同，且不断有用户上传新的视频文件
- 使用不同网络接入商的用户基数不一样，例如北京网通用户很多，而北京电信用户就相对少一些，使用DNS来实现LB不是最合适。

实时调度系统 – 对策

- 实时收集每个节点的带宽，动态调度每个播放请求，实时计算，不缓存
- 高峰期压力：10000+/s
- 丢弃数据库，丢弃memcached
- 数据横向Partition，共享内存多实例
- 数据、计算和策略分开

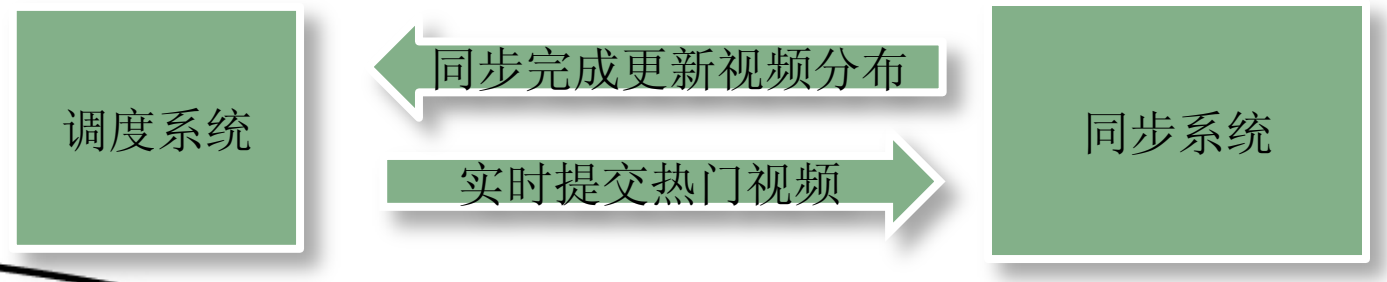
实时调度系统 – 系统结构

- **数据单元**负责提供调度时的数据。数据被加载到内存中，做横向partition，数据量庞大；需求最稳定，重启代价大
- **计算单元**直接接收用户端发来的调度请求，从数据单元获取数据，通过计算（过滤、排序）把播放链接返回给用户。计算负责，需求相对稳定；无状态，重启代价小
- **策略单元**向计算单元提供策略，不接收用户请求，计算量很小，但逻辑相对复杂且策略多变，重启无代价



调度与同步的实时反馈

- 一台机器比较合适的带宽输出是1.5G，对于码流是500k的视频，只能满足3000个同时在线。
- 因此对于一些热播剧，一个视频需要被分布在多台服务器上才能正常服务。
- 因此，调度和同步需要有实时交互



分享一些运维相关信息

- 节点管理平台，分布式监控平台，统一发布平台
- 不做raid，最大化每块硬盘IO
- 使用ext4文件系统，实测IO wait下降15%
- 使用deadline IO scheduler，IO能力提高5%

谢谢