360 HDFS下载平台介绍

唐会军

2011年12月2日

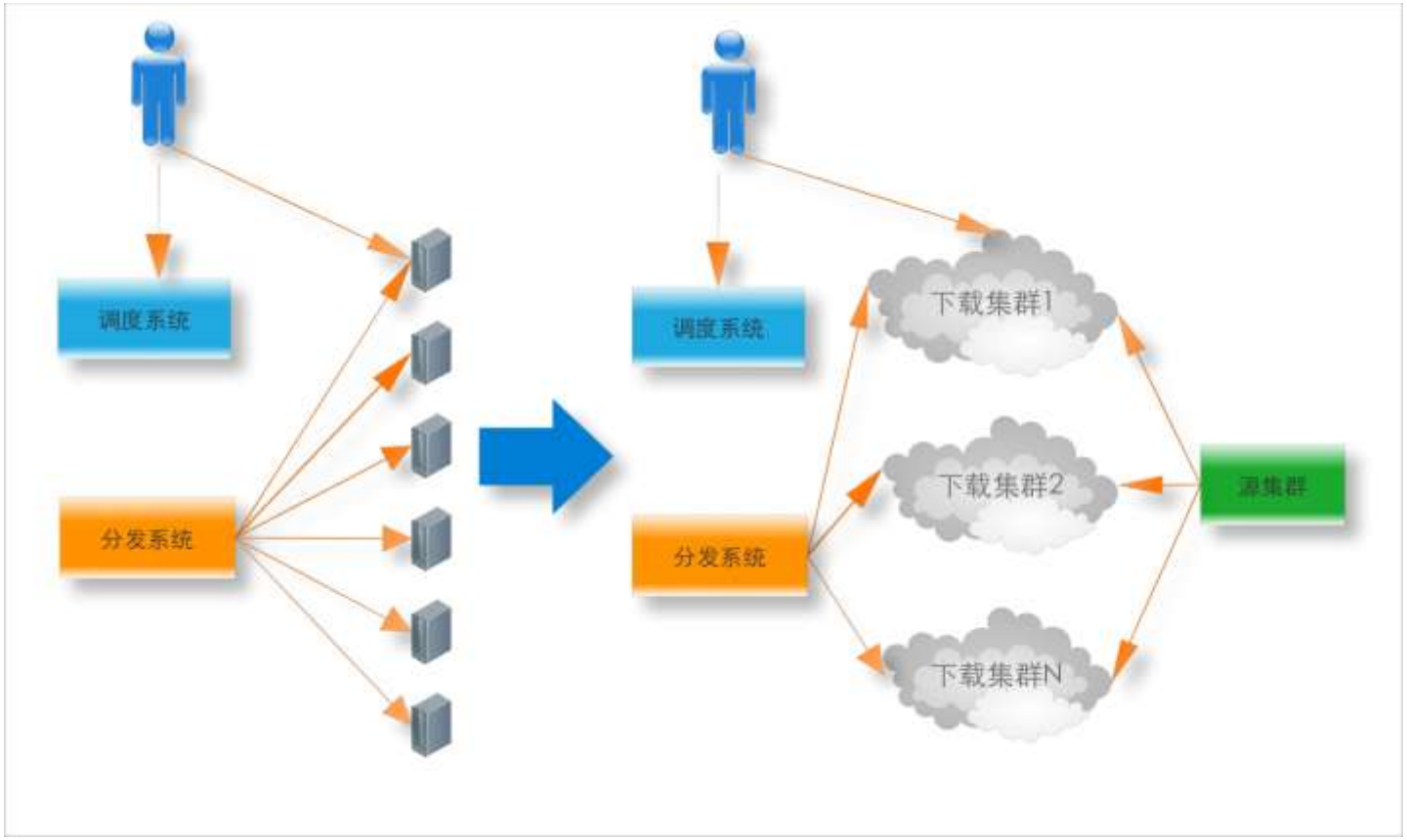
提纲

- 架构
- 优势
- 问题

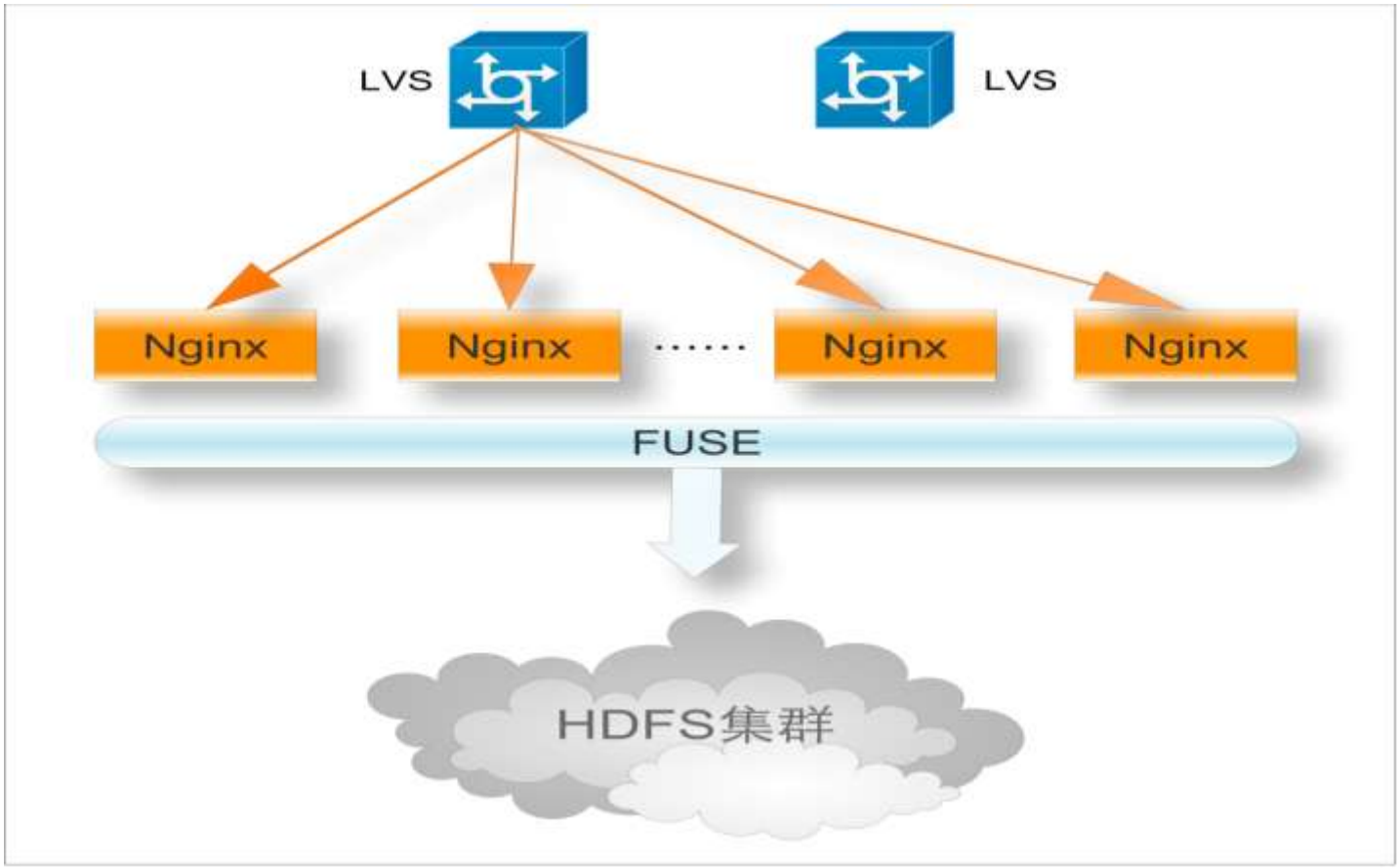
数据

- 下载平台相关数据：
  - 服务器总数接近 **1000** 台，流量接近 **300** Gb。
  - 部署在全国 接近 **100** 个IDC。

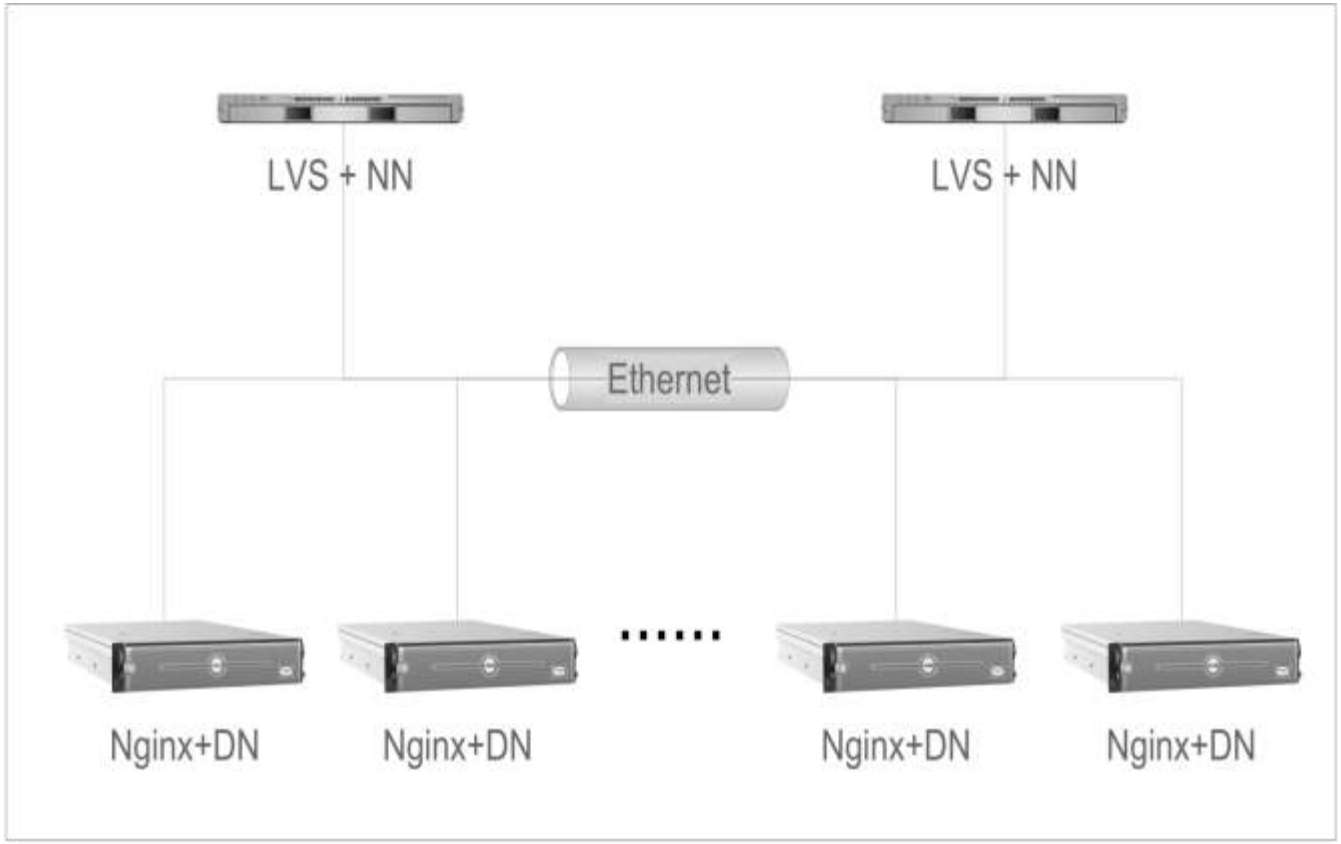
- 下载平台总体架构



- 下载集群架构图



- 下载集群部署架构



- 下载服务器配置
  - Nginx + Fuse + DataNode
  - 硬盘: 12 *SATA（不带Raid卡)
  - 网卡: 3块网卡（2个外网IP, 1个内网IP）

- DataNode
  - 副本数灵活调整（**3~8**）
  - 增大工作线程数(**1024**)，数据传输线程数(**4096**)

优势

- 发布和调度更简单
  - 发布和调度的基本单元由下载服务器变为HDFS集群
  - 集群数量远远小于服务器数量，极大简化了发布和调度的复杂度

优势

- 扩展性更好
  - 增加下载服务器（Nginx+DN）扩展单个集群的下载能力
  - 扩展过程不需要人工同步数据
  - 扩展过程对调度系统透明

优势

- 容错性更强
  - 全局：调度系统的健康检查机制确保集群宕后自动切走流量
  - 集群：LVS确保单台下载服务器宕机不影响集群的正常下载服务
  - 单机：HDFS 多数据副本机制确保硬盘的损坏不影响服务器的正常下载服务

优势

- IO性能更好
  - 硬盘较多时RAID卡成为瓶颈
  - 利用HDFS数据块分布机制达到充分使用多块磁盘I/O能力的目的
  - 目前线上12块硬盘IOPS峰值达到 1000左右, 而同等数量硬盘带RAID卡的服务器只能到500左右

问题

- 集群NameNode单点问题
  - NameNode宕机，导致集群不可用
- 解决办法
  - 目前主要利用集群之间的冗余，调度中心会和下载集群维持健康检查，一旦发现集群不可用，自动将流量调度到其它下载集群
  - 正在测试Facebook开源的avatar方案

问题

- 状态汇报影响性能
  - 文件个数增多后，块汇报，容量汇报极大影响性能和稳定性
- 解决办法
  - 将获取容量使用情况方法由 **du** 改为 **df**
  - 将块汇报中最费时间的获取所有块信息过程从心跳中剥离 (最新的unstable版本**1.1.0**已经有相应的patch)

总结

- 用HDFS提供在线下载服务是可行的
  - 目前在360使用了一年多，稳定性得到了考验
  - 极大降低了运维工作量，提高了服务的可靠性
  - 目标：通过更精细化的优化，将单台服务器的下载性能由目前的1Gb提高到2Gb


360
www.360.cn
Thanks!