大数据时代的网络架构

腾讯 杨志华

2013年8月

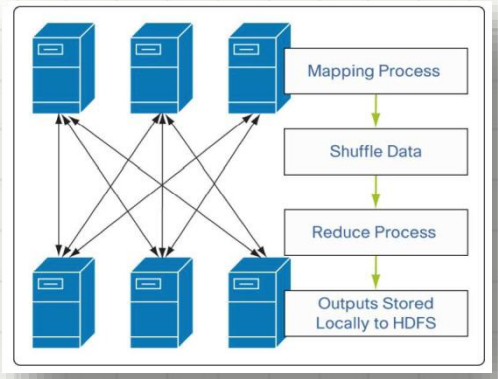
Agenda

- Bigdata对网络的挑战
- TCP Incast应对
- 低成本无阻塞网络架构
- SDN思路集群部署优化

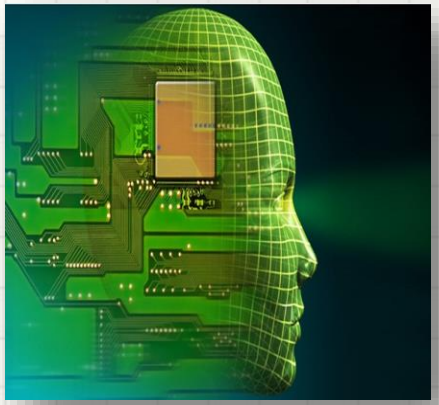
Three Things of Bigdata

SACC2013

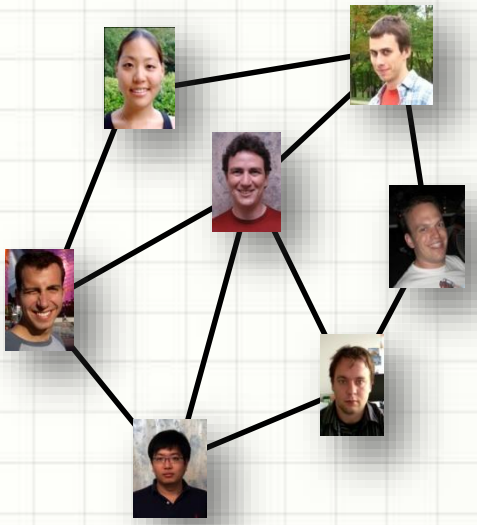
MapReduce



Deep Learning

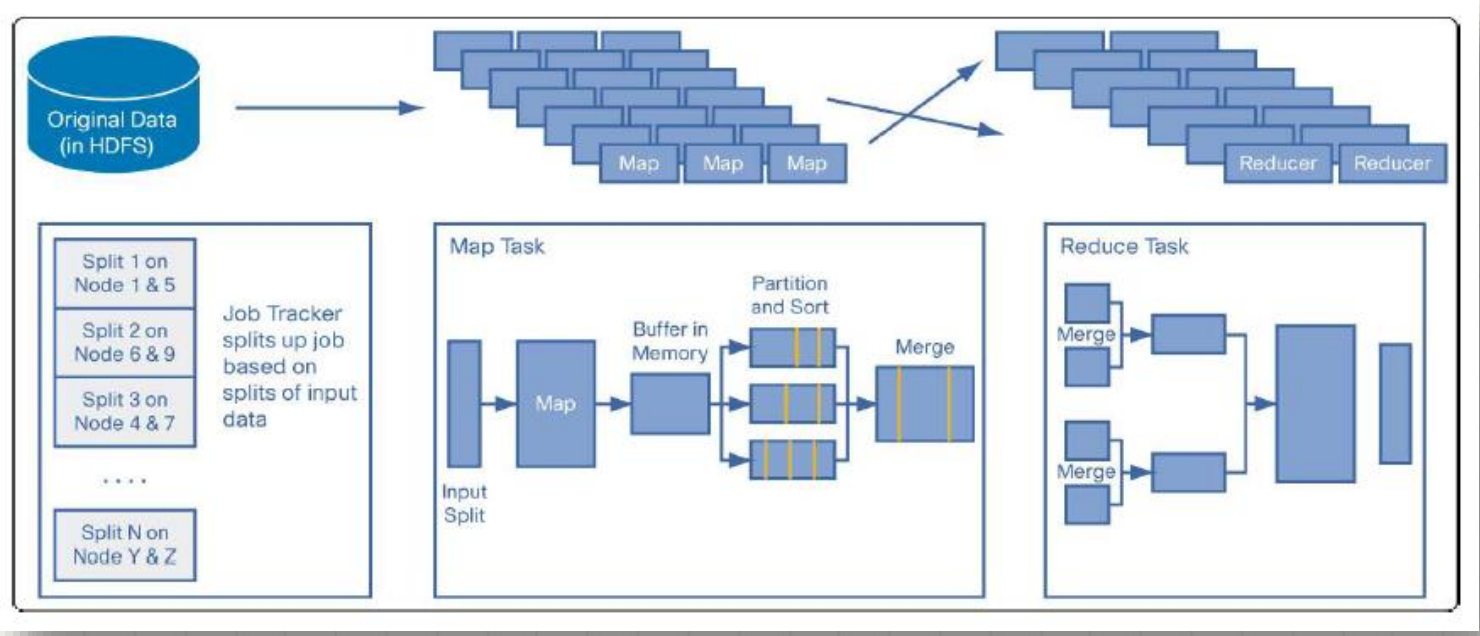


Graphlab



One Day of MapReduce

统计项	数据
集群规模	3000+
每日运行的作业数	25000+
每日运行的Worker数	150万+
日处理数据量	1PB+
日输出数据	200TB+

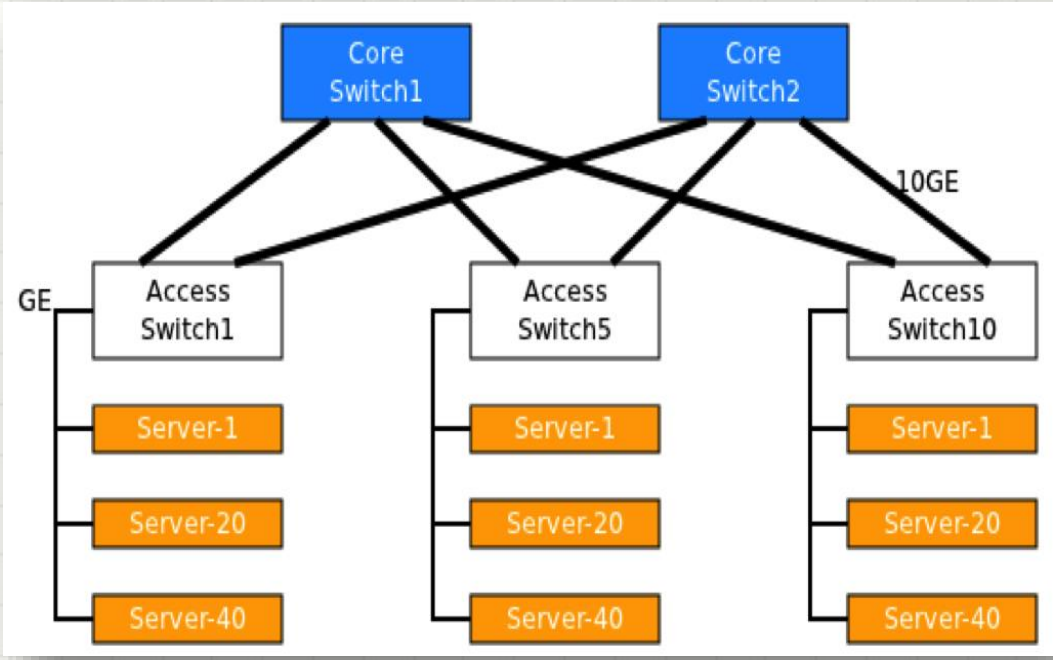


网络特点

- 离线计算
  - 无最终用户实时访问
  - 计算结果数据量大
- 高带宽集群通信
  - 轻易打满千兆 / 万兆
  - 集群规模几百、几千甚至更大
- 具备容错能力
  - 任务可调度，对少量节点故障不敏感
  - 数据冗余存储

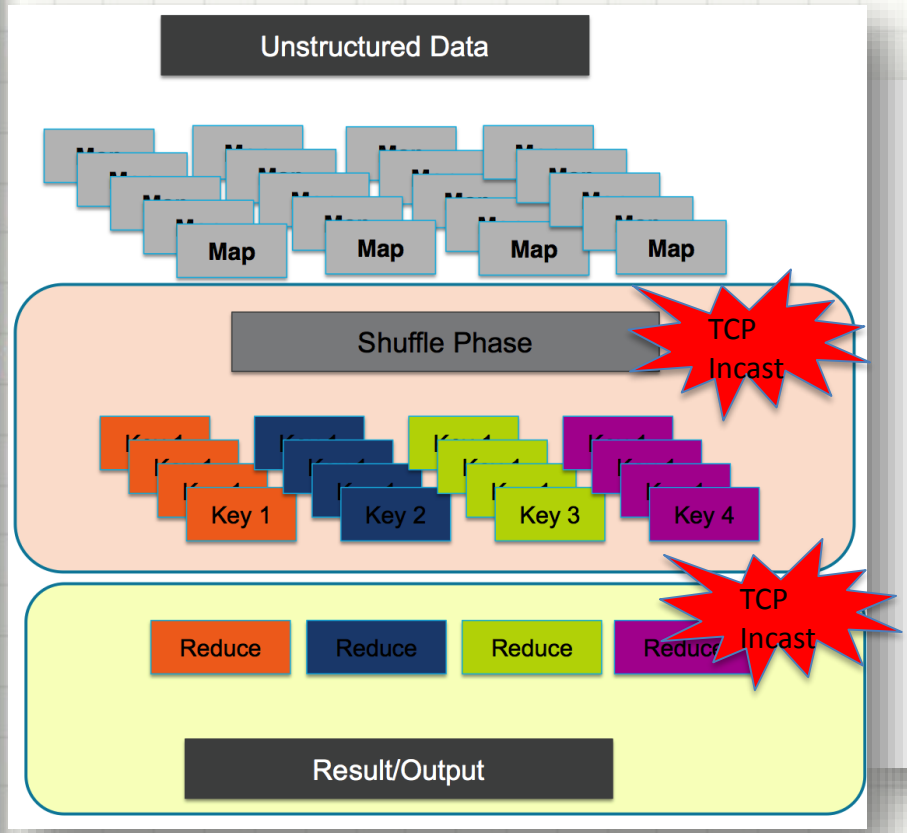
网络挑战

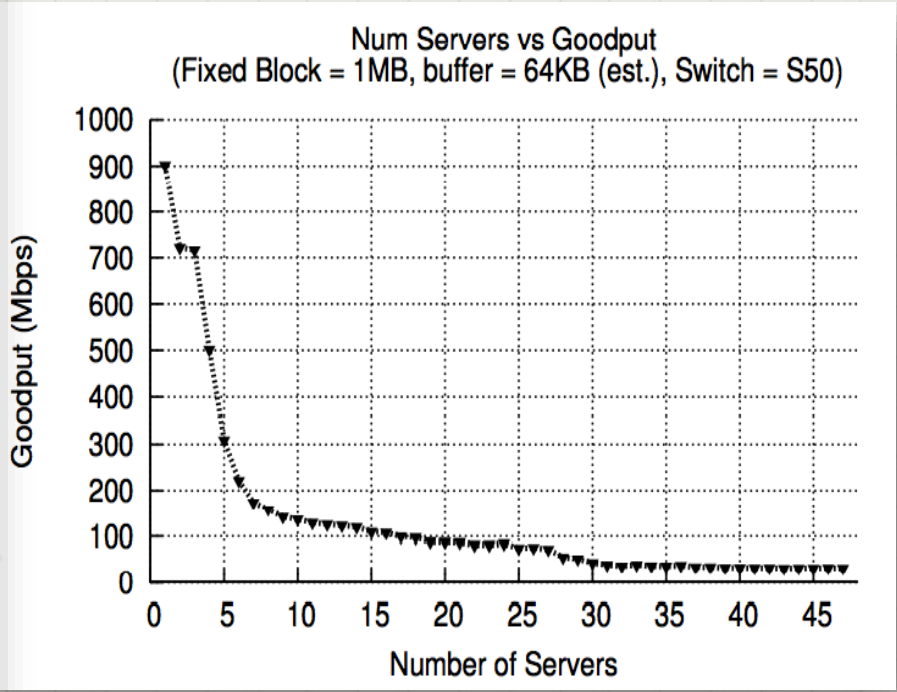
- TCP Incast
  - 几十、几百、几千打一
  - 传输性能优化
- 高带宽
  - 典型的网络重载业务
  - 持续大带宽通信
- 集群部署
  - 降低成本
  - 与其它业务混合部署
  - 减少突发对其它业务影响



TCP Incast场景

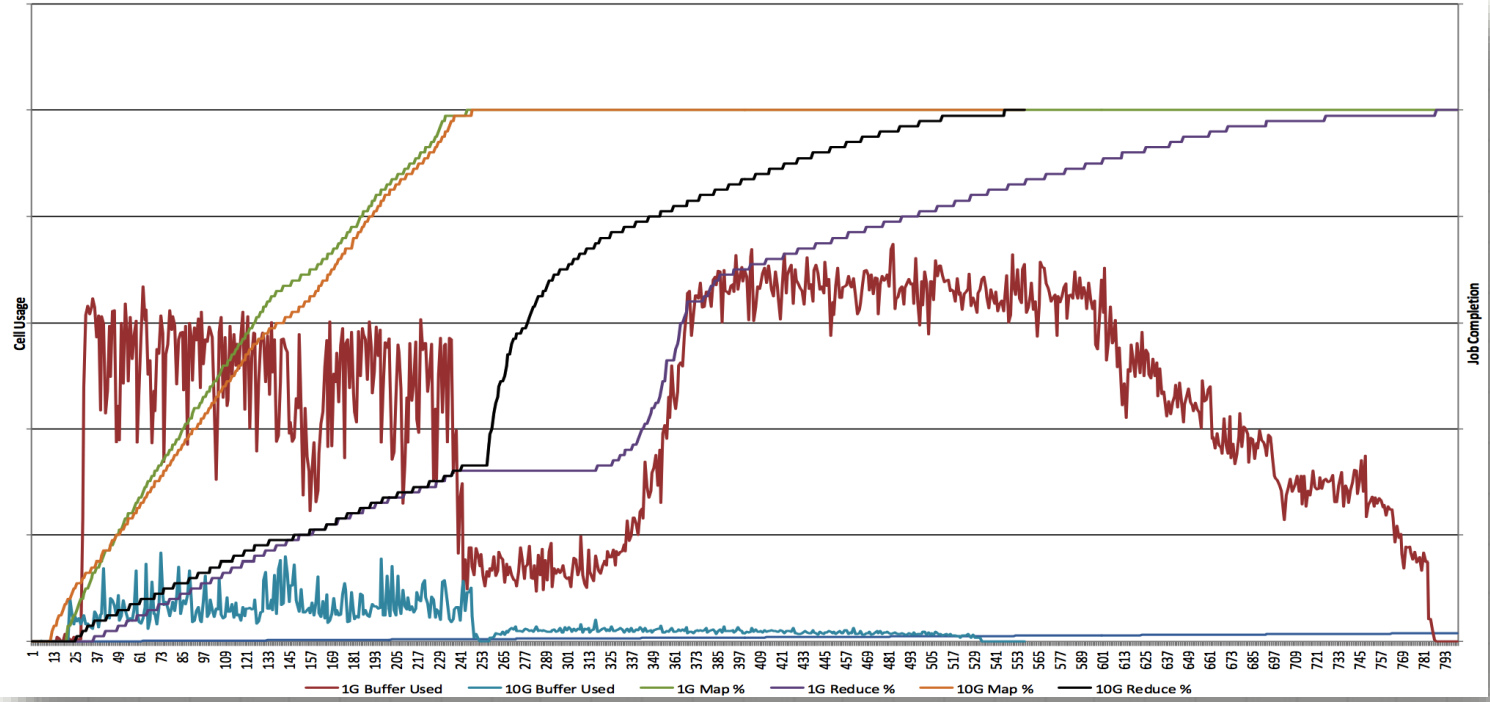
- 产生原因
  - 多打一的大量突发导致网络端口丢包
  - 丢包导致TCP超时重传





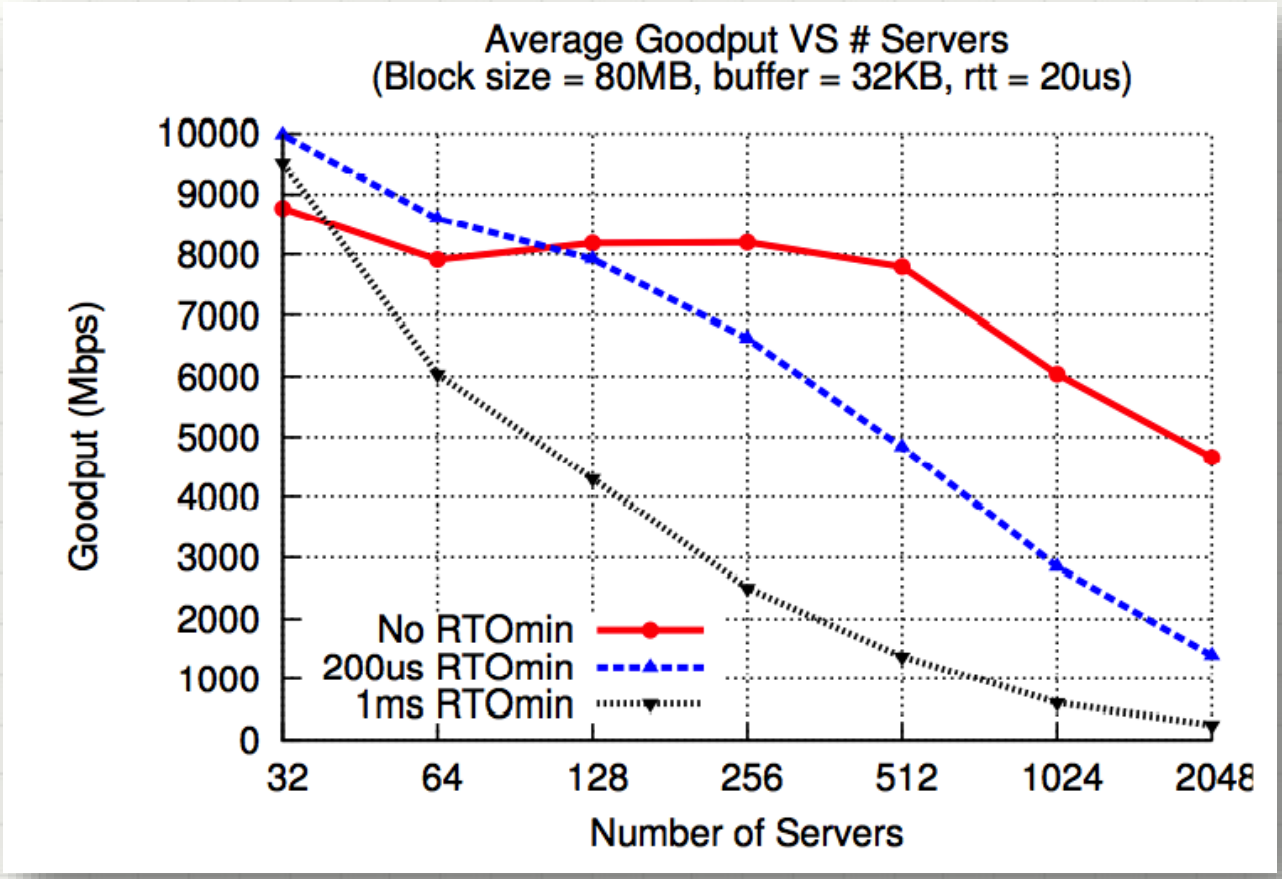
交换机缓存：越大越好？

- 交换机缓存设计与加速比等各方面相关
- 合理的缓存可以减少丢包但对突发的应对有限
- 1G升级至10G可以更有效的应对突发



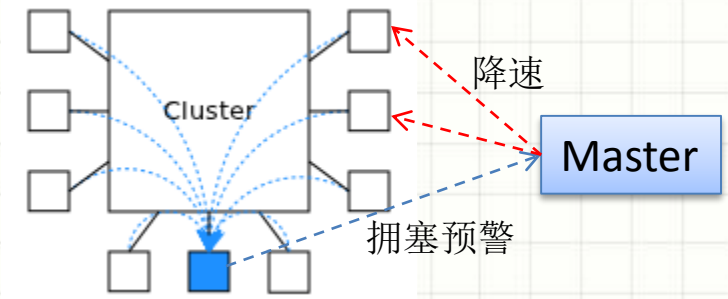
系统TCP调优

- 拥塞窗口值优化、Qick ACK、SACK等
- 减小TCP RTOmin。参考资料：Safe and Effective Fine-grained TCP Retransmissions for Data-center Communication



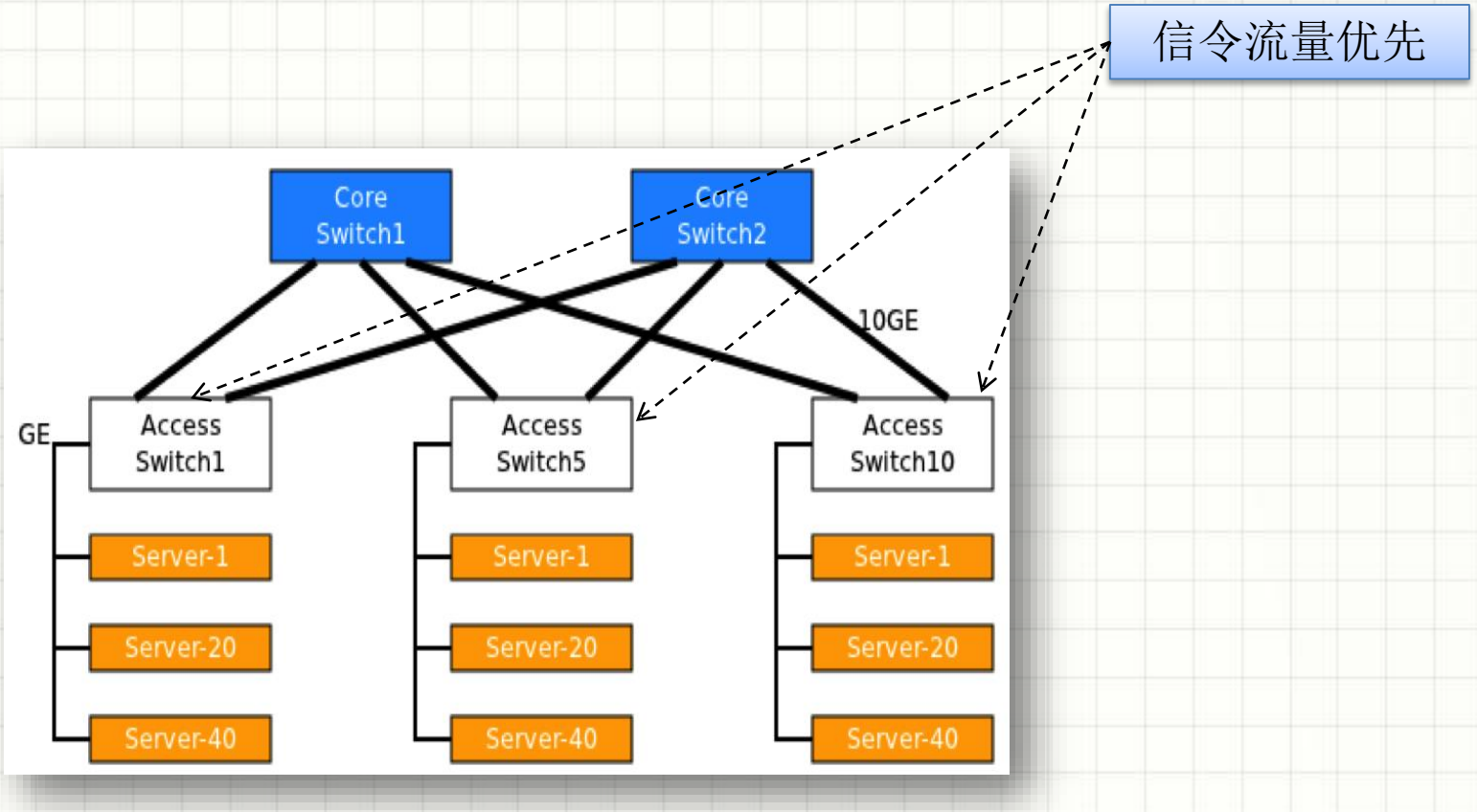
任务调度优化

- 任务管理
  - 任务分配考虑数据存储分布，减少数据拉取流量
  - 流水线shuffle、reduce
- 监控网络，负压反馈
  - 接收方监控网络利用率
  - 达到阈值通过信令反馈至发送方降低流量
  - 提高网络利用率，缩短任务时间



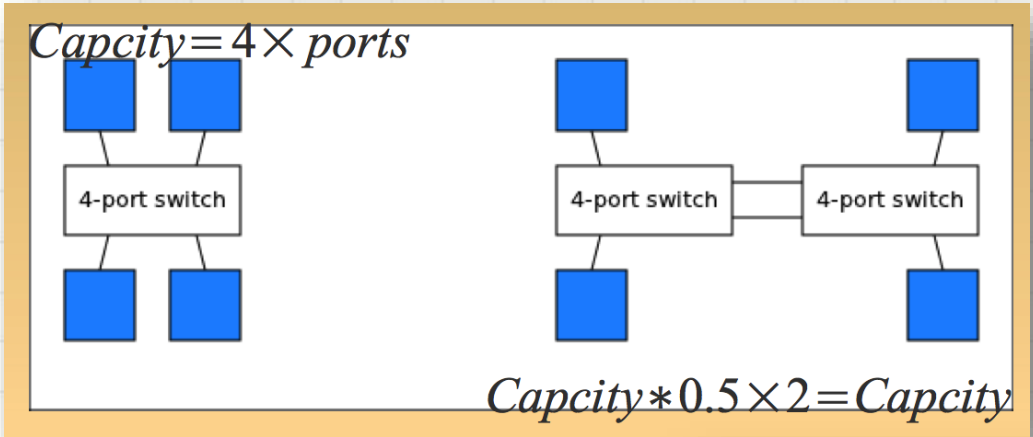
数据中心内部网络QoS

- 接入层交换机QoS确保任务调度的信令流量优先传输
- 应用对信令流量进行DSCP标记



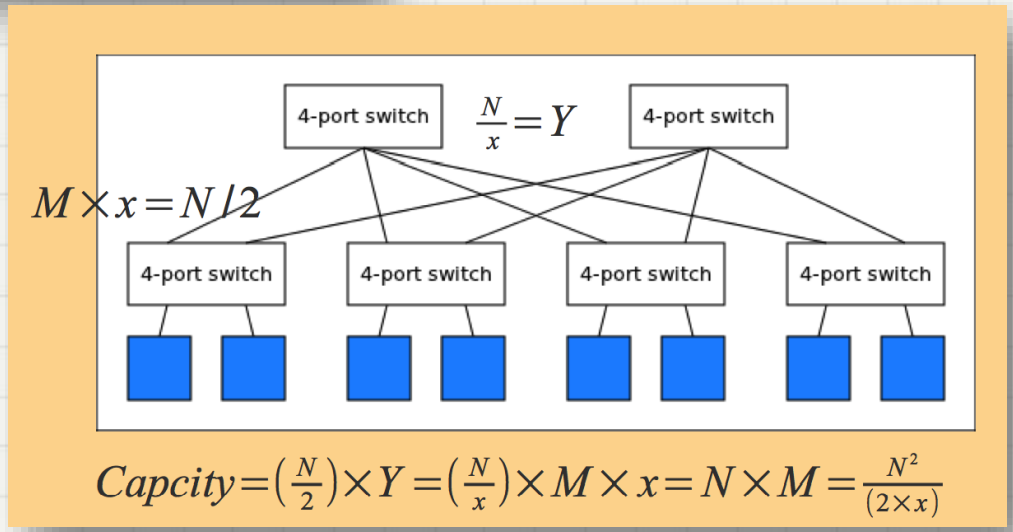
低成本的无阻塞Clos网络架构 SACC2013

- 使用固定端口交换机组建大规模无阻塞网络



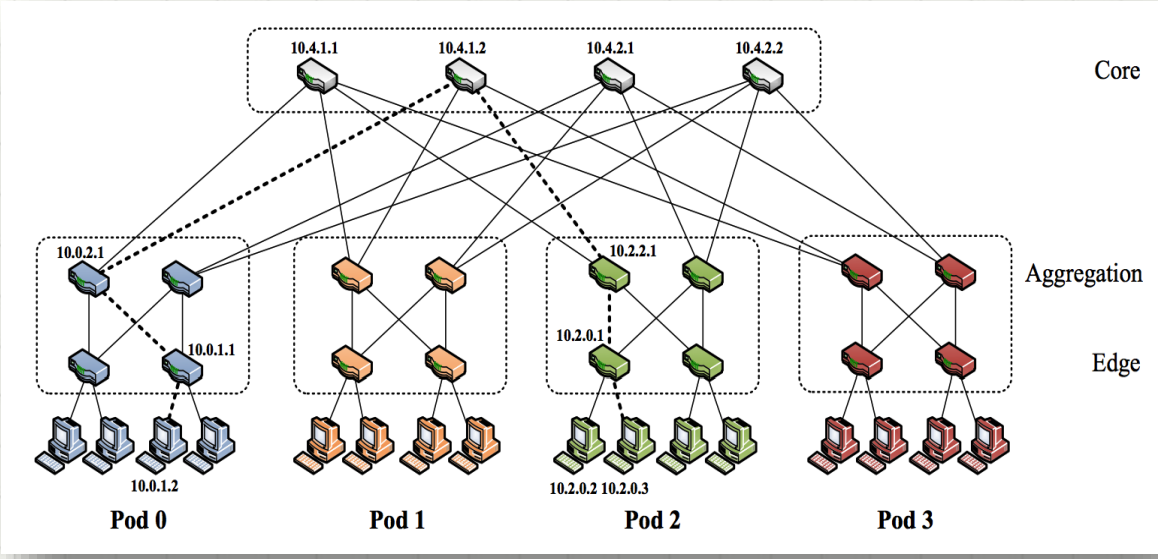
- 以4端口无阻塞交换机为例
- 2台4端口交换机通过无阻塞互联结果还是4端口无阻塞

- 6台4端口交换机通过无阻塞互联结果是8端口无阻塞



低成本的无阻塞Clos网络架构（续）

- 72台48端口千兆交换机搭建1152端口千兆无阻塞网络
- 96台64端口万兆交换机搭建2048端口万兆无阻塞网络
- Google：2008年搭建3级Clos网络，20480端口千兆无阻塞网络

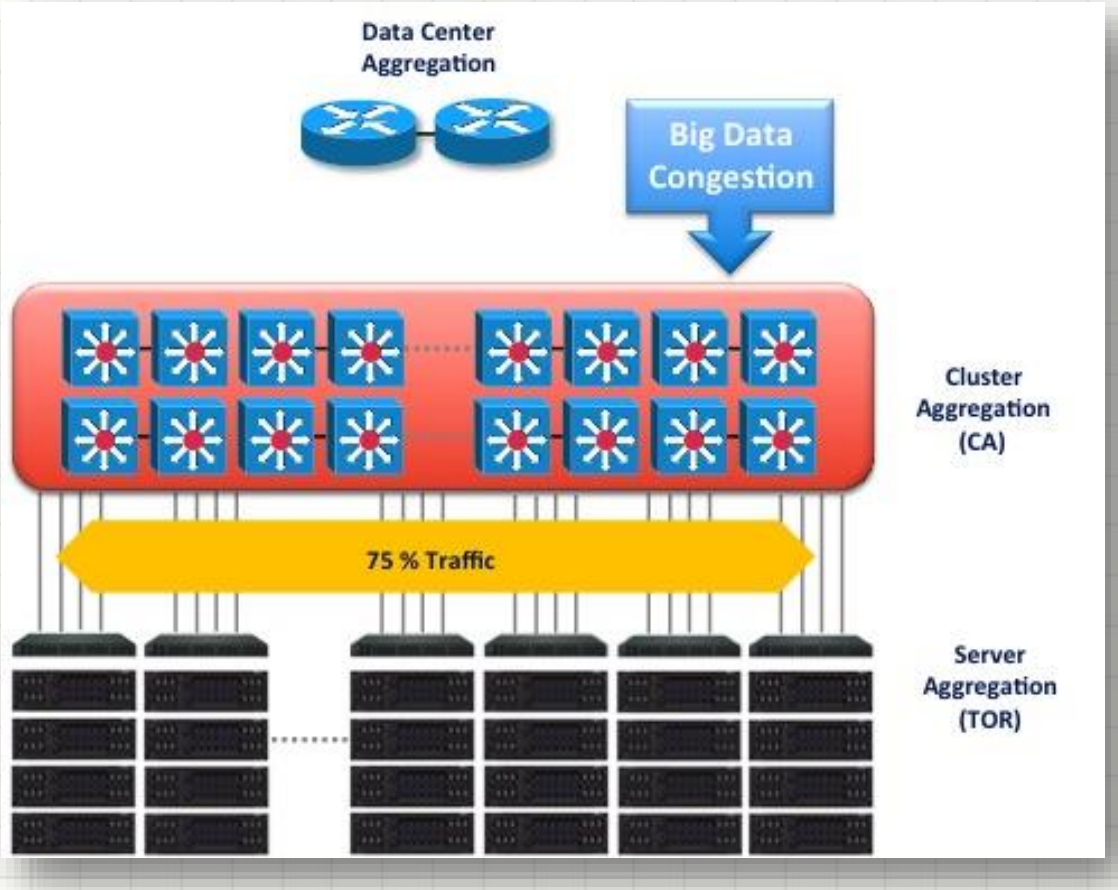


- 优势
  - 低成本,1/2~1/6
  - 机架电力要求低
  - 可无限扩展
- 挑战
  - 大量布线
  - 大量管理节点
  - ECMP数量

参考资料：A Scalable, Commodity Data Center Network Architecture

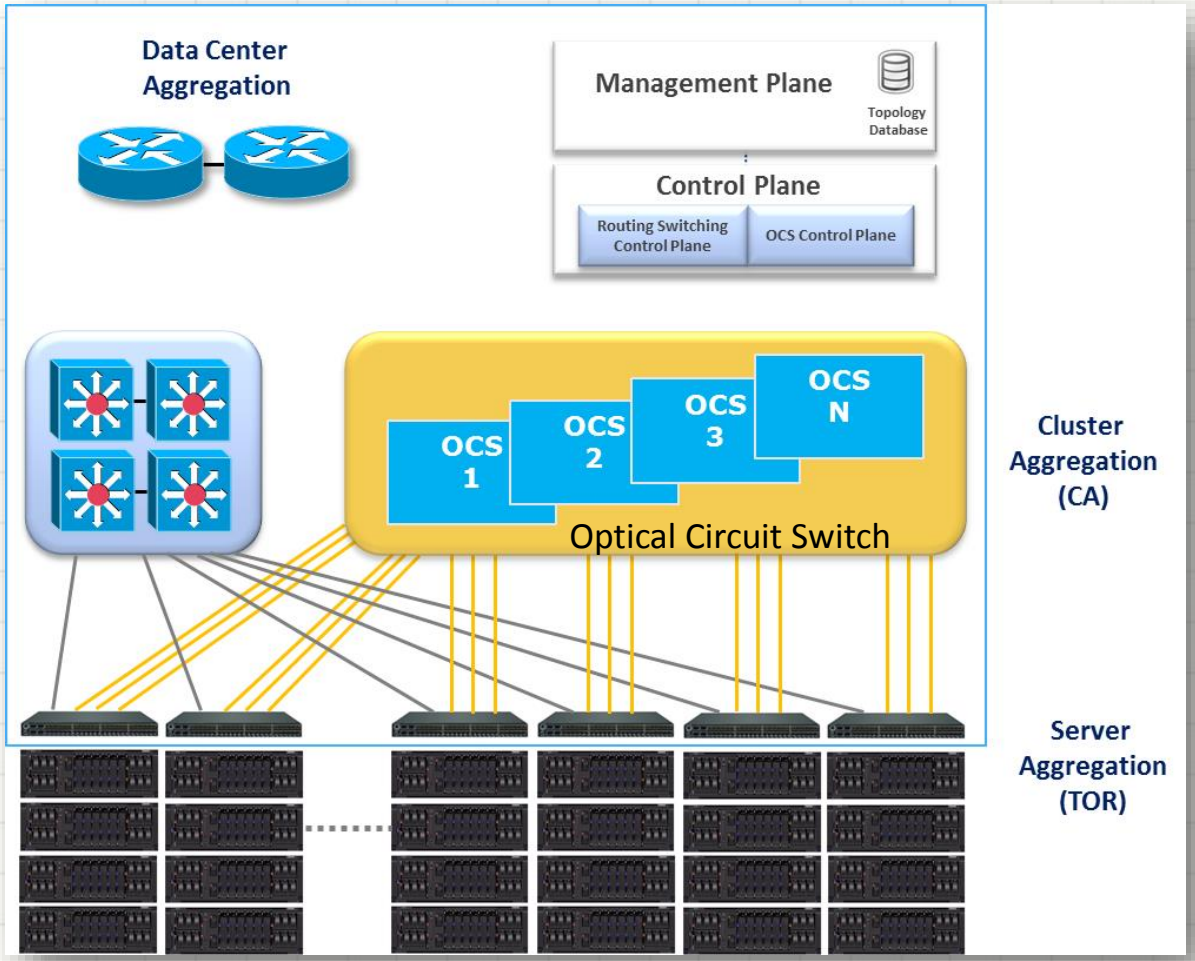
集群部署的考量点

- 普通网络难以支持Bigdata集群
- 大带宽占用冲击其它业务



SDN思路光调度集群网络

- 传统网络核心提供南北向通道
- 光网络按需提供集群高带宽东西向互联通道



Key Takeaways

- Bigdata网络的特点是离线、大突发、高带宽
- TCP Incast主要通过任务调度优化来解决，TCP调优为辅，网络QoS可用于确保信令
- Clos网络架构可搭建低成本无阻塞网络
- SDN思路的光网络调度方案可低成本解决集群与其它业务混合部署、适应普通网络场景

Tencent 腾讯 | 一切以用户价值为依归

THE END