China · Beijing

百度第三代Spider背后的万亿量级实时数据处理系统

颜世光

百度 架构师

ArchSummit 全球架构师峰会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ

InfoQ

促进软件开发领域知识与创新的传播

关注InfoQ官方微信

及时获取ArchSummit

大会演讲视频信息

QCon

全球软件开发大会 [北京站]

2017年4月16-18日 北京·国家会议中心

咨询热线：010-64738142

ArchSummit

全球架构师峰会 2016 [深圳站]

2017年7月7－8日 深圳·华侨城洲际酒店

咨询热线：010-89880682

个人介绍

- 颜世光, 百度搜索基础架构技术负责人
- 代表作品
  - 百度第三代Spider系统
  - 百度文件系统 BFS
  - 集群调度系统 Galaxy
  - 海量实时数据库 Tera
- 个人主页&Blog&微信
  - https://github.com/bluebore
  - http://bluebore.cn

大纲

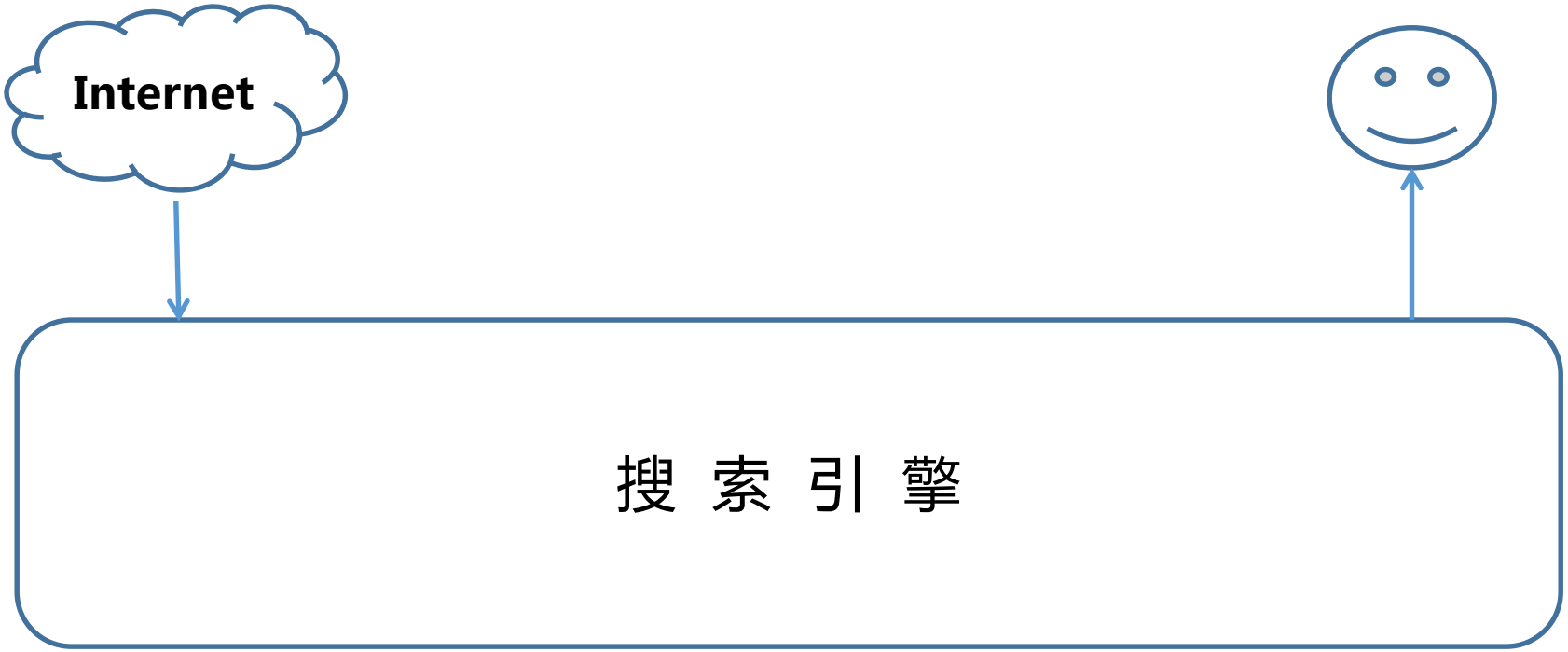
搜索引擎与Spider3.0

Tera的模型与架构

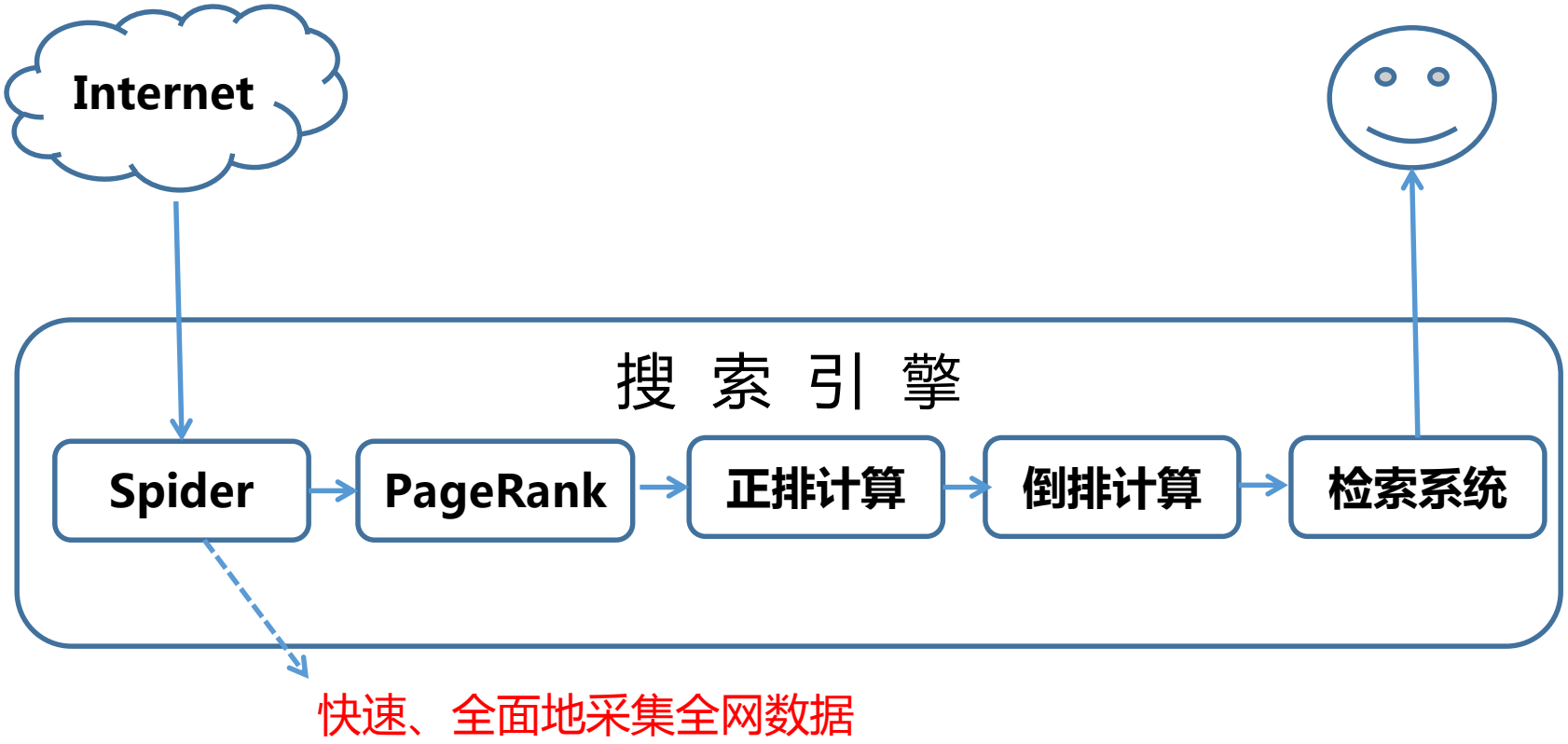
系统构建中的经验与教训

未来工作

互联网与搜索引擎



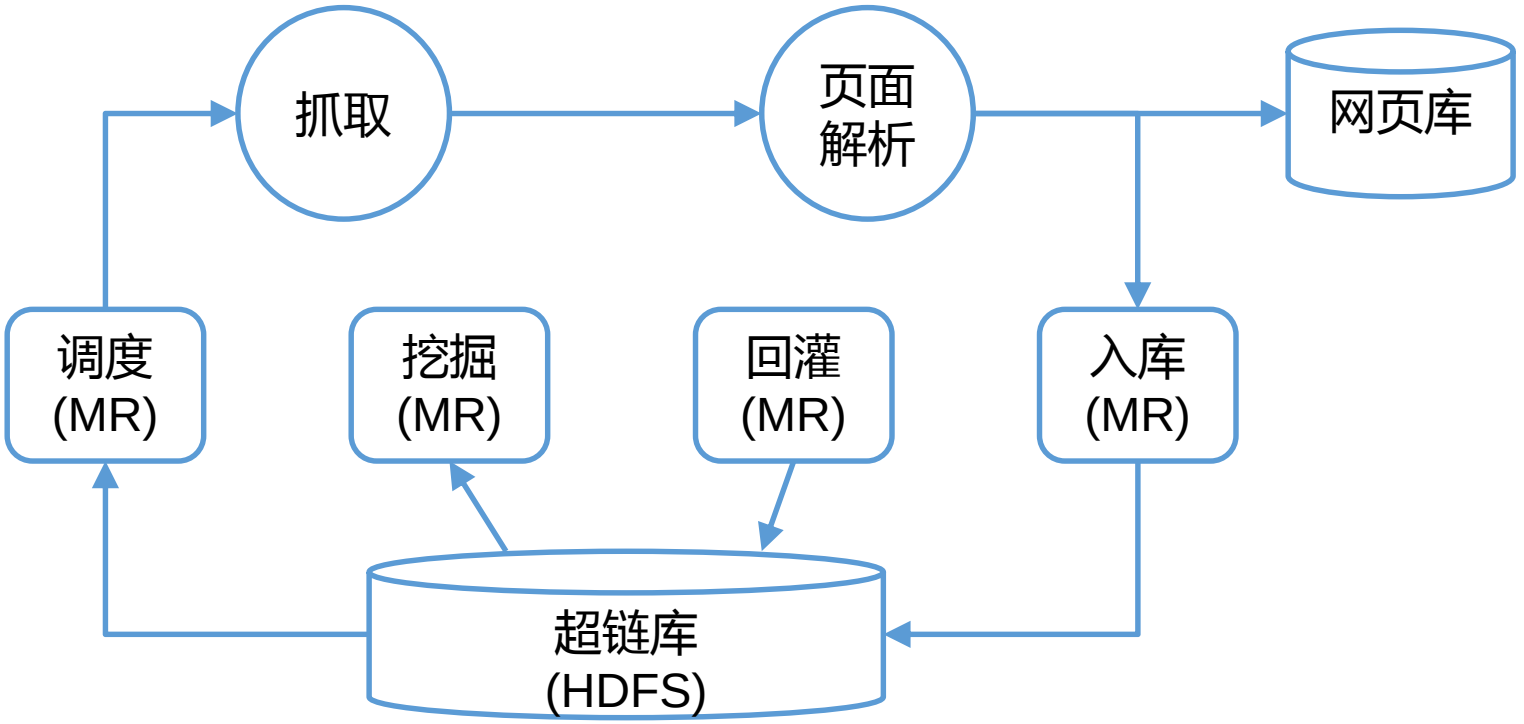
搜索引擎与Spider



中文互联网与百度Spider

- 网页总数: 100万亿
- 有价值网页: 10万亿
- 每天新增: 100亿
- 超链接数: 120 条/网页
- Spider每天处理提链: 100亿 * 120 = **1.2万亿条**

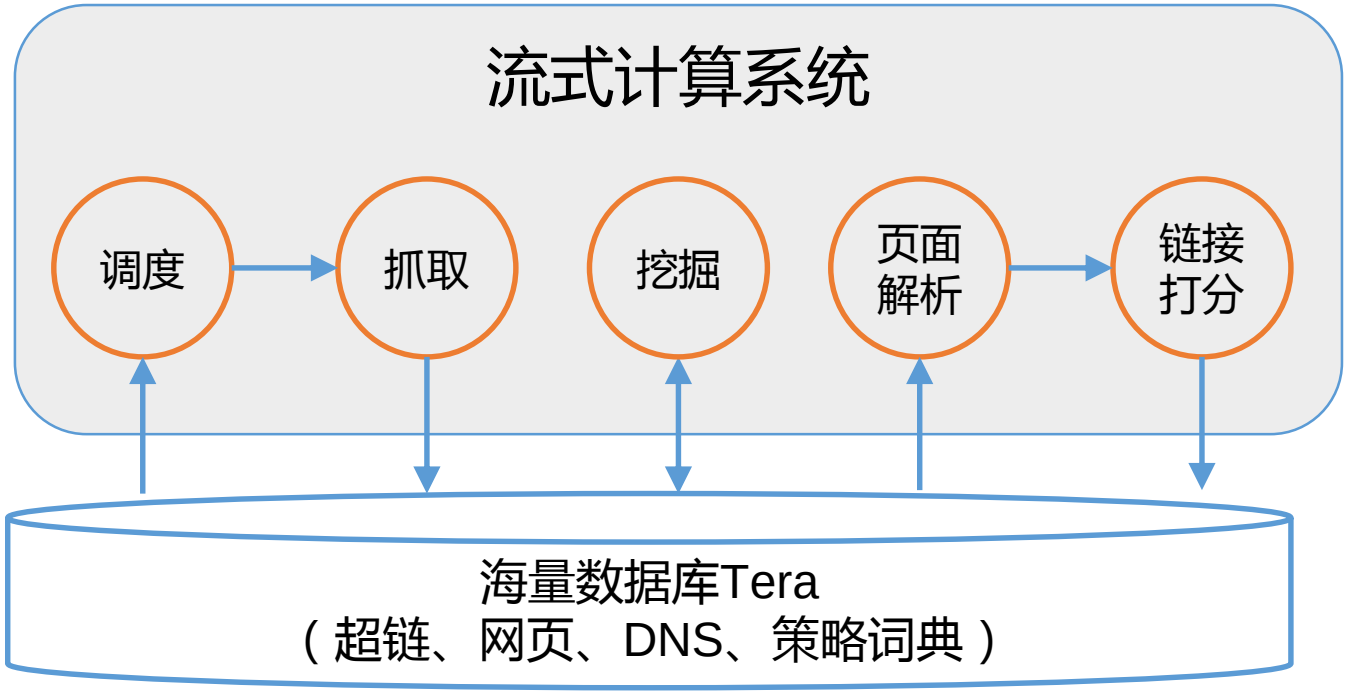
Hadoop时代的百度Spider



Hadoop的问题

- 线性扩展问题
  - 1000亿链接处理 -> 500台服务器
  - 10万亿链接处理 -> 5万台
  - 解决：必须增量处理
- 时效性问题
  - 近10轮MR过程，耗时两天
  - 解决：必须流式处理

百度第三代Spider



实时处理的核心

- 数据是本质
  - 来源是数据
  - 产出也是数据
  - 中间状态
- 一条新链接的价值谁说了算?
  - 站点&路径深度
  - 前链&锚文本
- 一张网页变化
  - 触发上百条链接属性更新

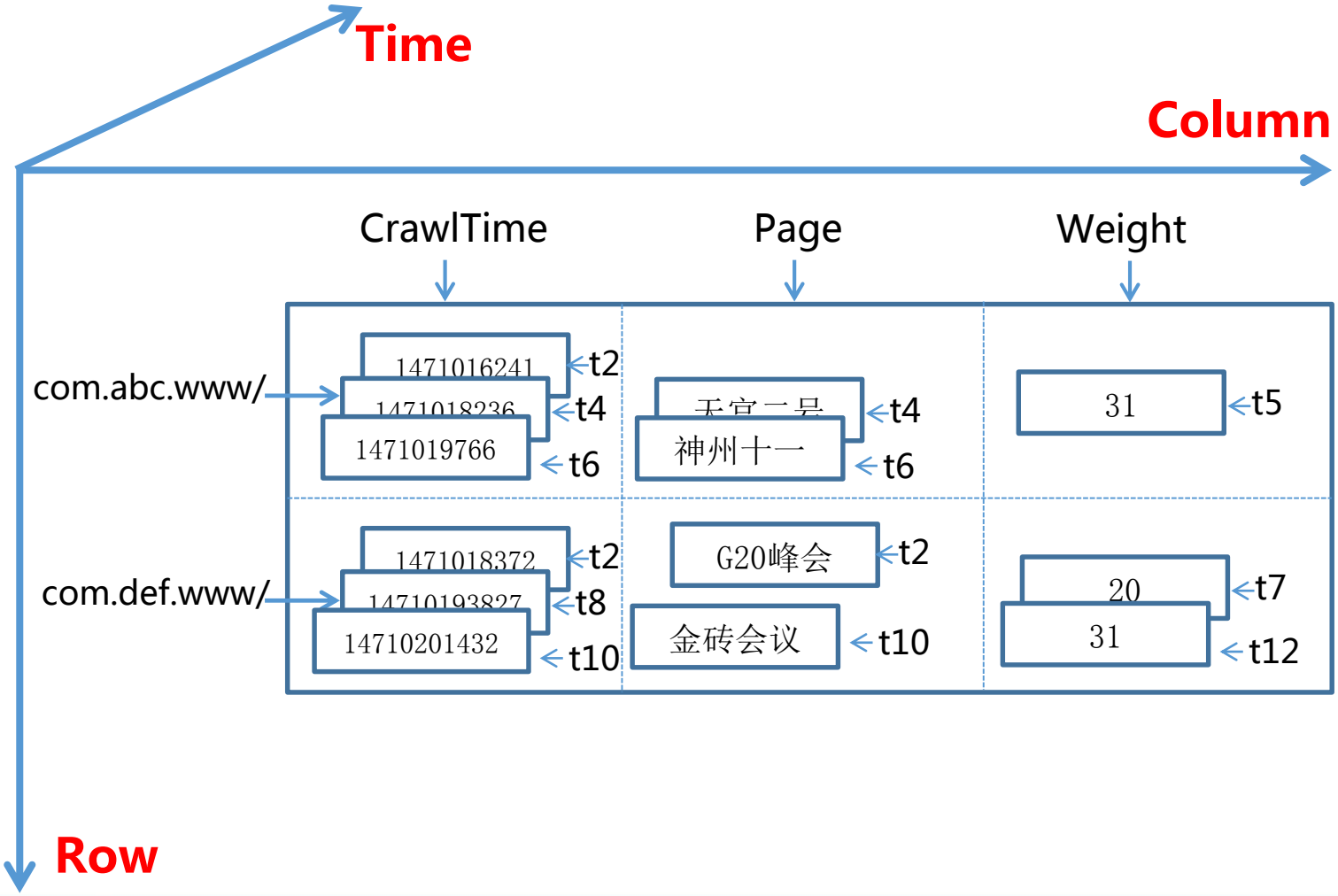
Spider3.0的实时数据处理

- 全量数据 ~10万亿条 ~100PB
  - 每一条随时都可能更新
- 每天新抓网页 100亿
  - 触发1万亿条链接更新
- 每秒属性更新 ~1亿次
  - 随机读&随机写
- 全局调度
  - 站点&主域压力受控
  - 虚拟主机运营商压力受控

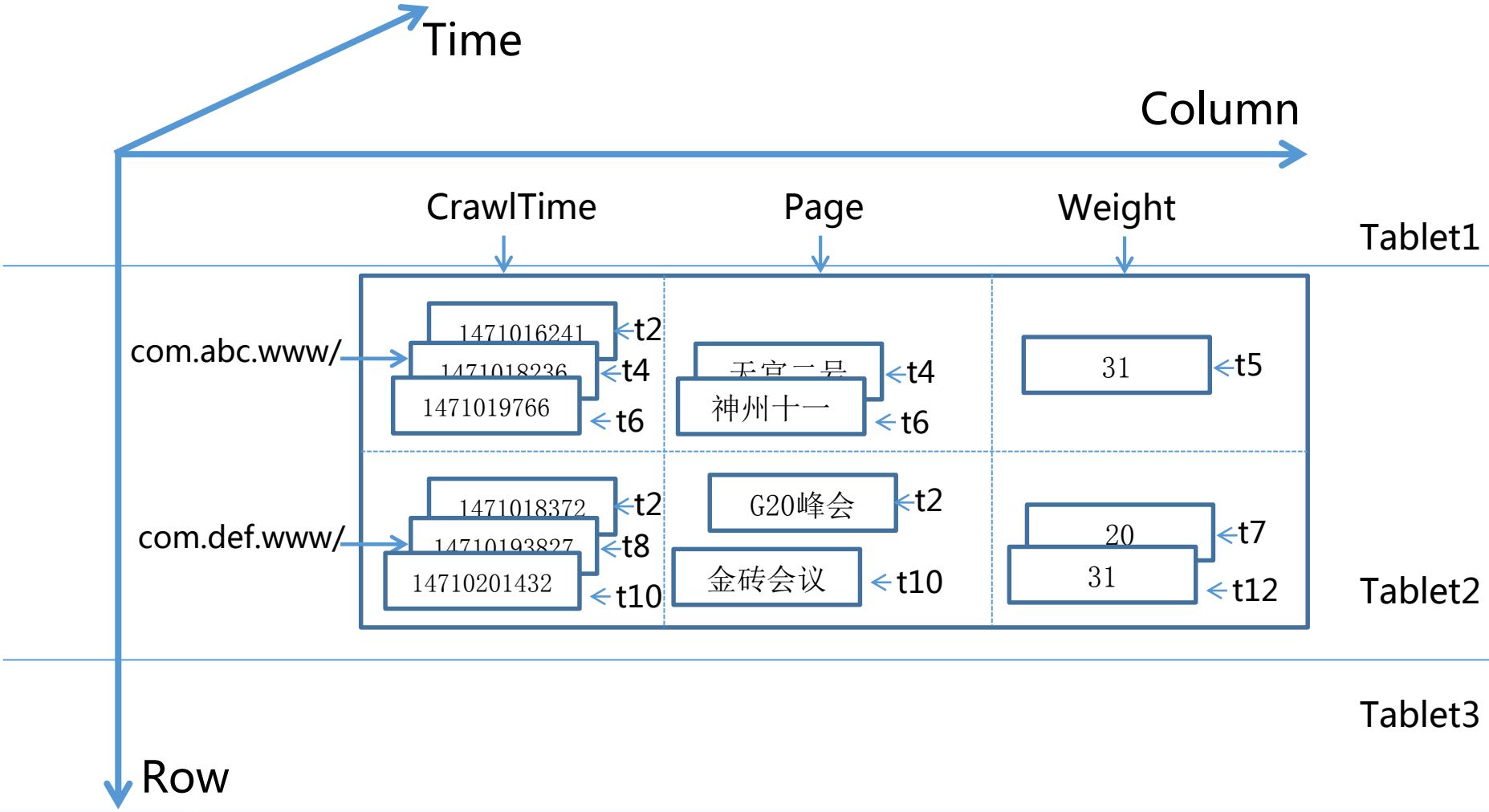
我们的解决方案

- 海量实时数据库Tera
  - 分布式、可扩展
    - **万亿**记录数，**百PB**容量，**亿级QPS**读写
  - 全局有序表
    - 支持区间访问，方便统计
  - 自动负载均衡
    - 互联网热点频发，业务迭代迅速
  - 多版本、表格快照
    - 历史数据分析、业务数据回滚
  - 其他特性
    - 列存储、分布式事务

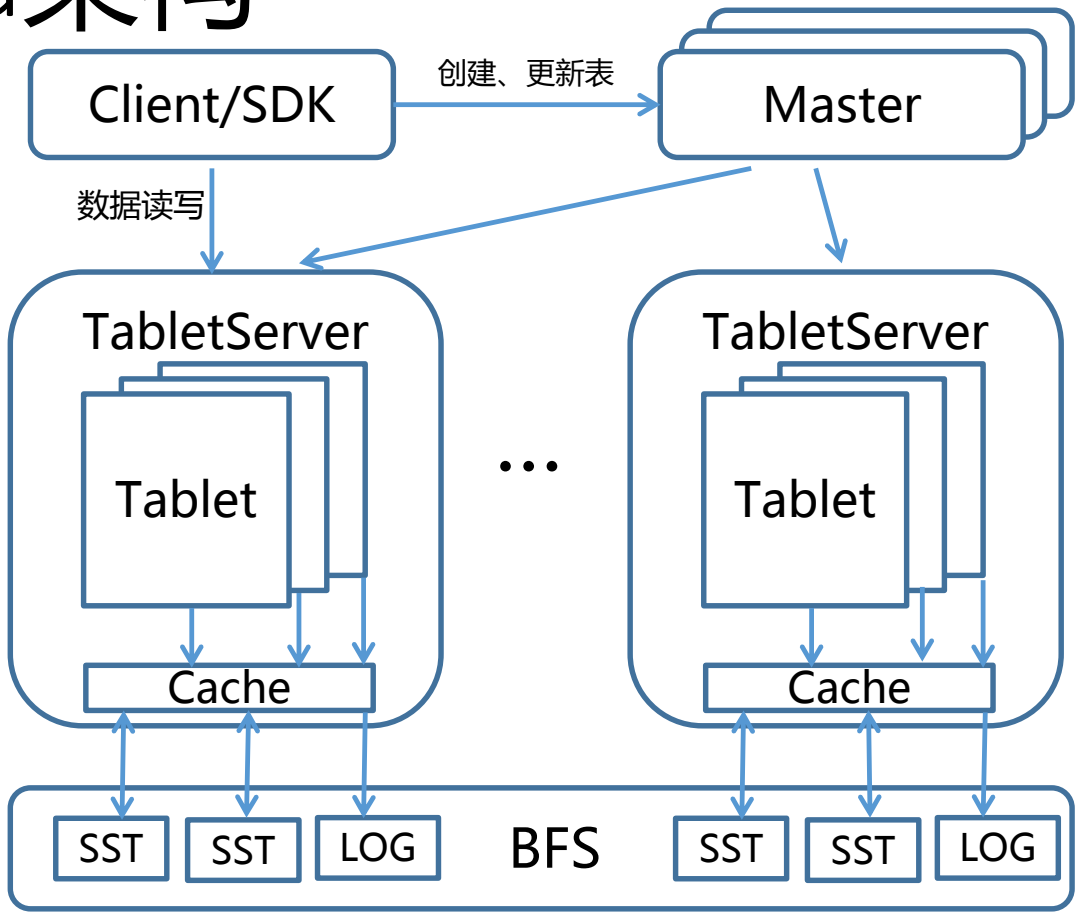
Tera的表是三维的


Time
Column
Row
CrawlTime
Page
Weight
com.abc.www/
1471016241
t2
1471018236
t4
1471019766
t6
t4
神州十一
t6
31
t5
com.def.www/
1471018372
t2
14710193827
t8
14710201432
t10
G20峰会
t2
金砖会议
t10
20
t7
31
t12

按行切分成多区间(Tablet)



Tera架构



- 先写内存再写Log，文件全部持久化在分布式文件系统上。
- LOG: Write-ahead log，正常情况下只写不读，用于容灾。
- SST: 内存Dump或Compaction产生的静态文件，只读不改。

Tera给我们带来了什么?

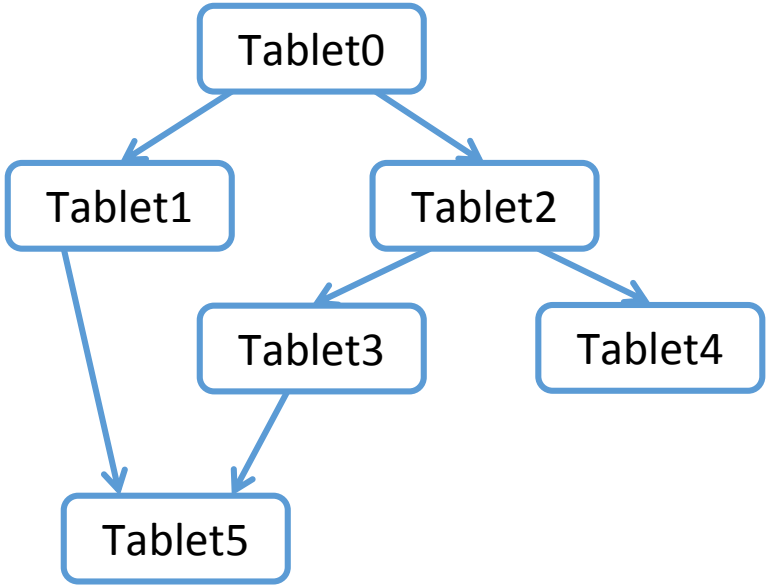
- 海量数据随时、随处可用
  - PB级的内存，统一的地址空间
  - 百PB级存储，不用担心持久化
  - 亿级QPS的吞吐承载
  - 毫秒级的延迟

对比HBase

- 相同点
  - Bigtable数据模型
  - 开源
- 不同点
  - 可用性
    - 解决了区间热点问题
    - 99.9% -> 99.99%
  - 性能、延迟
    - C++实现，没有GC问题
    - Locality Group支持
  - 扩展性
    - 数百台->数千台

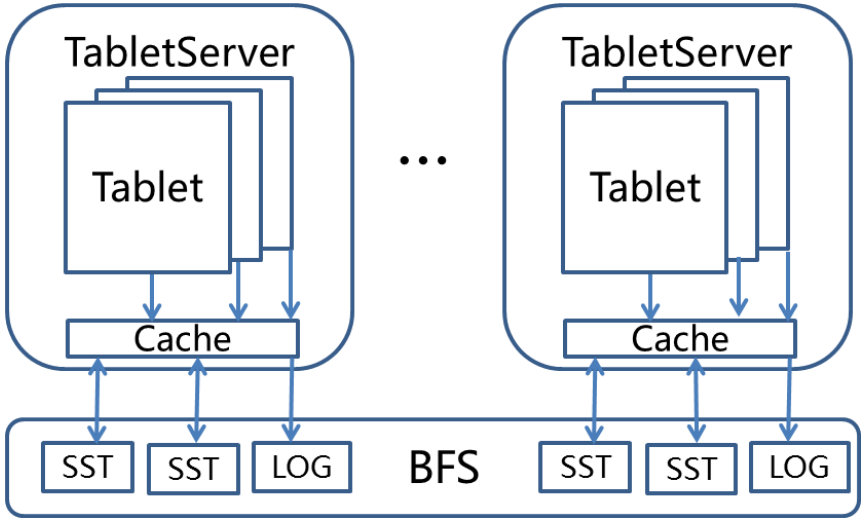
快速负载均衡

- 分裂快
  - <50ms
  - 通过文件引用实现
- 敢分裂
  - 很好地处理碎片问题
  - 热点过后，快速合并回来



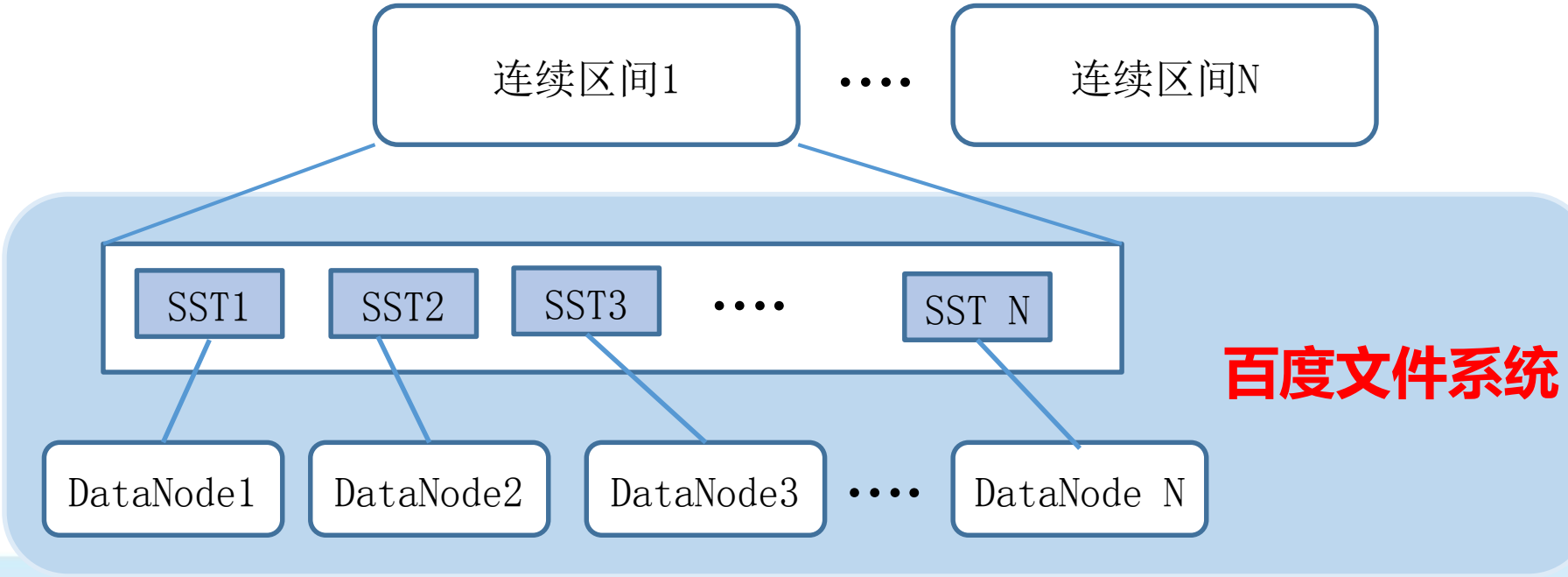
能快速合并，才敢分裂

- 区间快速迁移
  - <50ms
  - Powered by BFS
- 区间快速合并
  - 仅元数据变更
  - 代价小, 时间短(200ms)
  - 全自动
  - 无人工干预



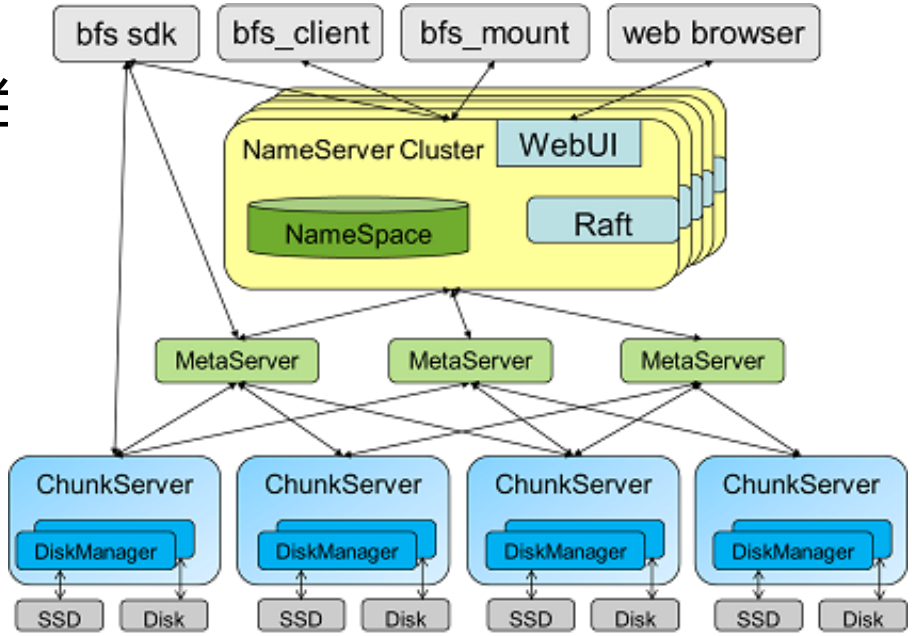
热点问题根本解决

- 分布式文件系统
  - 表面上：实现了快速分裂与迁移
  - 本质是：天然将请求打散到数千节点



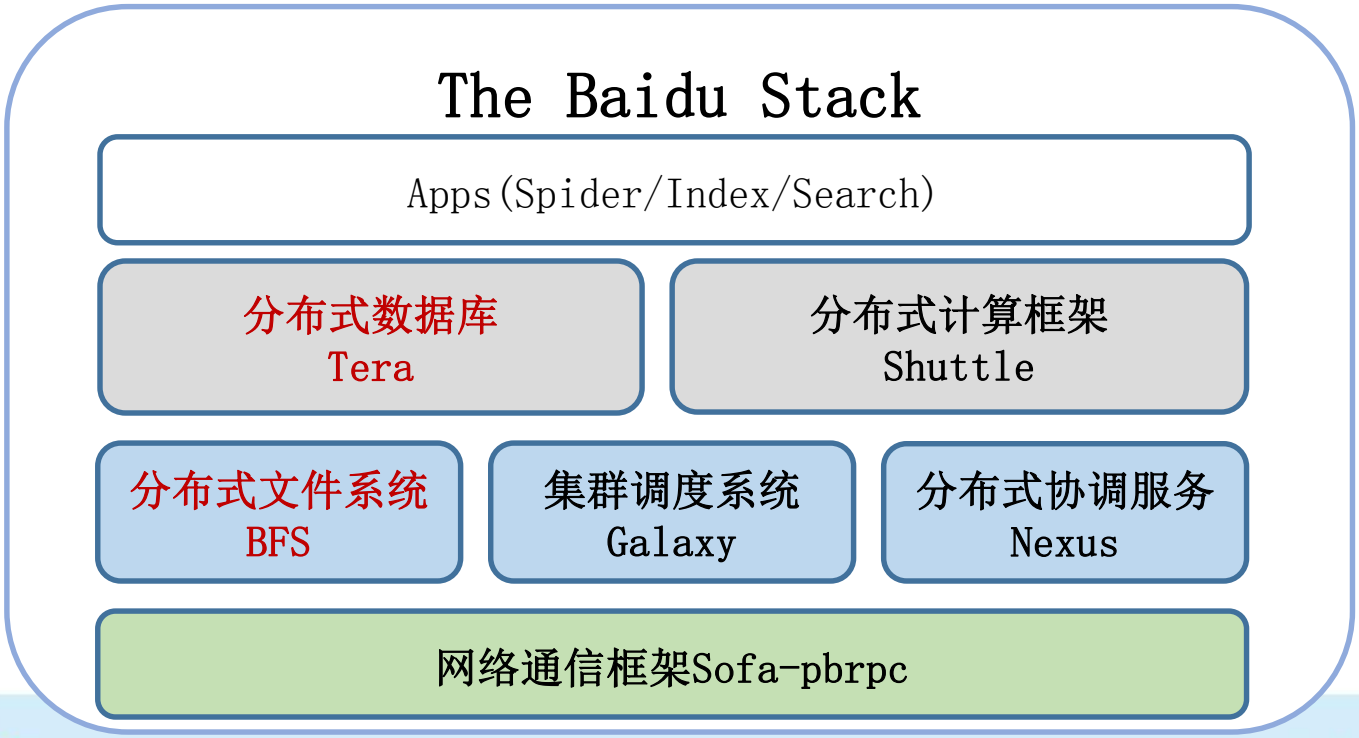
面向实时应用的百度文件系统

- 元数据可用性
  - 无NameNode单点
  - 基于Raft的分布式集群
- 文件可用性
  - 多数据中心副本放置
  - 快速副本恢复
- 高吞吐、低延迟
  - C++实现
  - 针对读写长尾优化
  - 单机 1.1GB/S读写吞吐



工业实践－分层设计

- 分工、复用
  - 问题最好解决一次
  - 一处解决多处受益



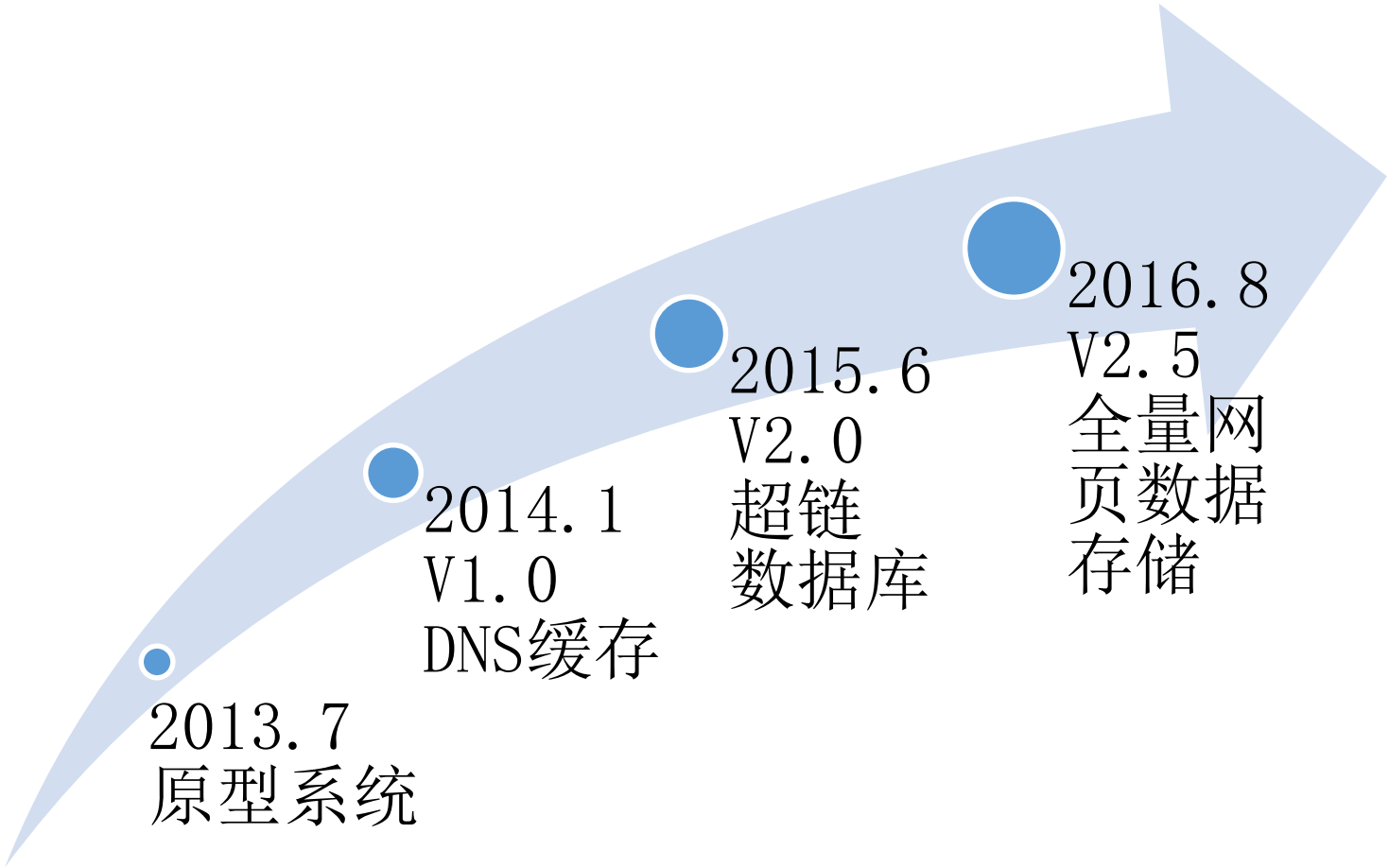
工业实践 – 可用性设计

- 硬件&软件故障不可避免
  - 假设有MTBF是30年的机器
  - 搭建一个1万台的集群
  - 每1~2天坏一台
- 降低故障恢复时间
  - 可用性 = (总时间 - 故障数 * 恢复时间) / 总时间
  - HBase 几分钟
  - Tera 几秒钟

工业实践 – 低延迟设计

- Backup Requests
  - 2ms后发送备份读请求到第二个副本
  - 如果一个被响应了，Cancel掉另外一个
  - 99.9分位延迟降低80%
- 慎用自动GC的语言
  - 实时处理, 大量小请求,频繁触发STW
  - 服务无响应
  - 不必要的failover

Tera在百度发展


2013.7
原型系统
2014.1
V1.0
DNS缓存
2015.6
V2.0
超链
数据库
2016.8
V2.5
全量网
页数据
存储

Tera在百度的应用

场景	描述	数据规模	天级读写
DNS信息存储	站点IP、Robots	~10TB，~100亿条记录	~100亿次
超链数据库	全网超链接数据	~10PB，~10万亿条记录，每条记录数百列	~1万亿次
网页数据库	全网网页数据	~100PB，~1万亿条记录，每条记录数百列	~1万亿次
内容池	用户&内容数据	很小，但快速增长	很小

未来工作 - 走出百度，走向社区

- github.com/baidu/tera
- 开发和Code Review都在GitHub上
- 来自社区的贡献已经在驱动百度搜索

欢迎加入Tera开发者交流

THANKS

ArchSummit
全 球 架 构 师 峰 会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ