58同城的分布式存储架构实践

58同城技术中心架构部
徐振华

SACC2012 2012-09-13

SACC

58同城.com

SACC

分布式存储基础知识

如何设计分布式存储架构

58的分布式存储实践

SACC

引入

Google Dremel 原理 – 如何能3秒分析1PB

磁盘的顺序读速度在100MB/S上下，那么在1S内处理1TB数据，意味着至少需要有1万个磁盘的并发读!

Dropbox声称，每天要接受2亿次上传,用户总数达到了2500万,估值会在50亿美元到100亿美元之间。

某云存储服务商官方声明,称因为机房的一台物理机本地磁盘损坏，导致个别用户数据丢失。

SACC

分布式存储基础知识

存储基础

分布式基础(存储相关)

SACC

存储基础

存储的理论和应用

存储系统的目标

SACC2012

SPAT

58同城.com

SACC

存储理论和应用

I/O五分钟法则 (局部性原理)

应用:无处不在的缓存;

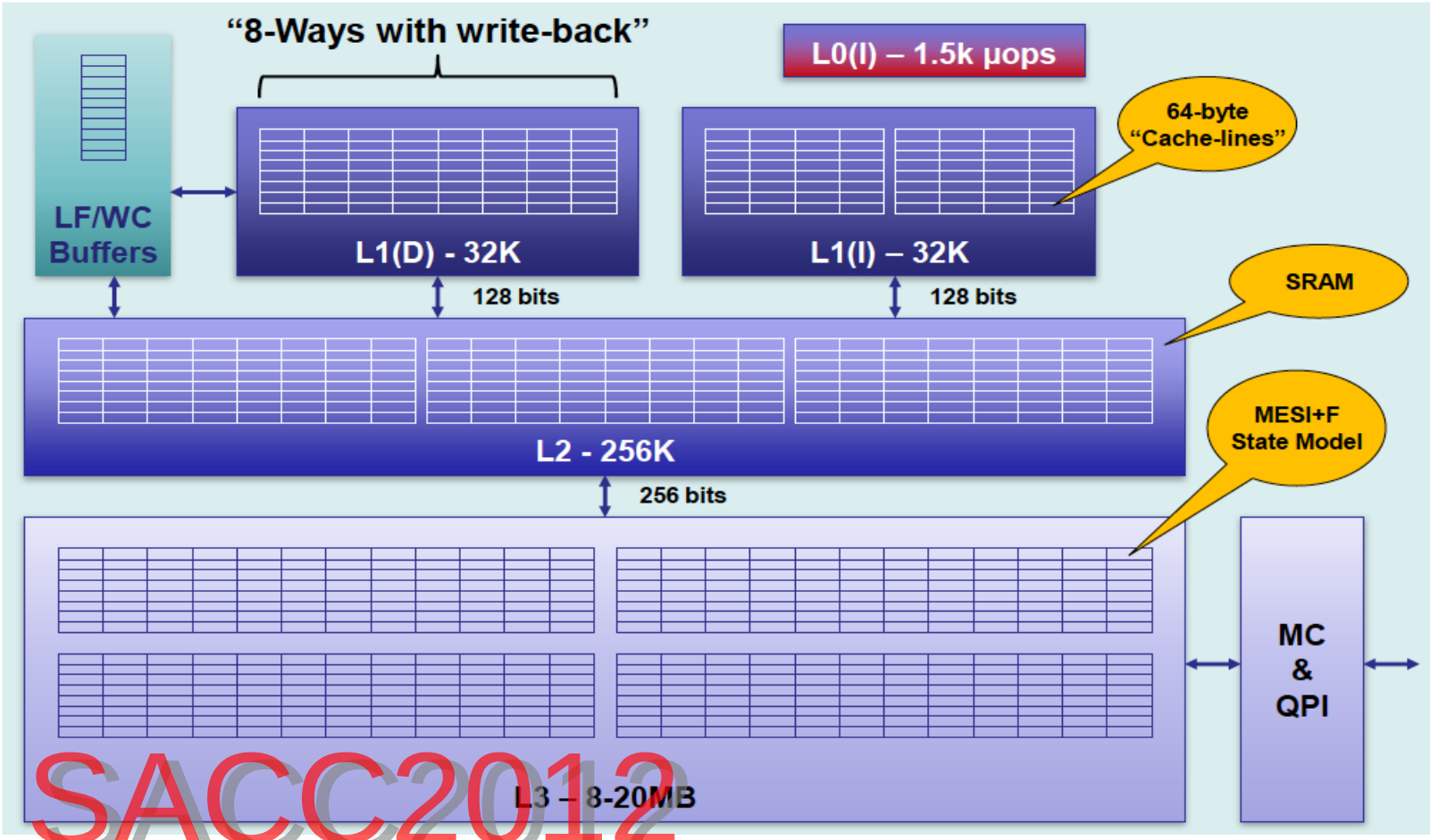
应用:热数据放在比较快的介质上;

Amdahl定律和Gustafson定律 , 摩尔定律

应用:提升系统性能

SACC

cpu缓存



Linux文件系统缓存

SATA II 7200 RPM IOPS: ~90
SAS15K RPM IOPS: ~180
Intel X25-M IOPS: ~8600
PCIe 2.0x8 IOPS~220,000 (4KB)

扇区 512,
内存页4k ,
磁盘块大小 4k
mtu 1500

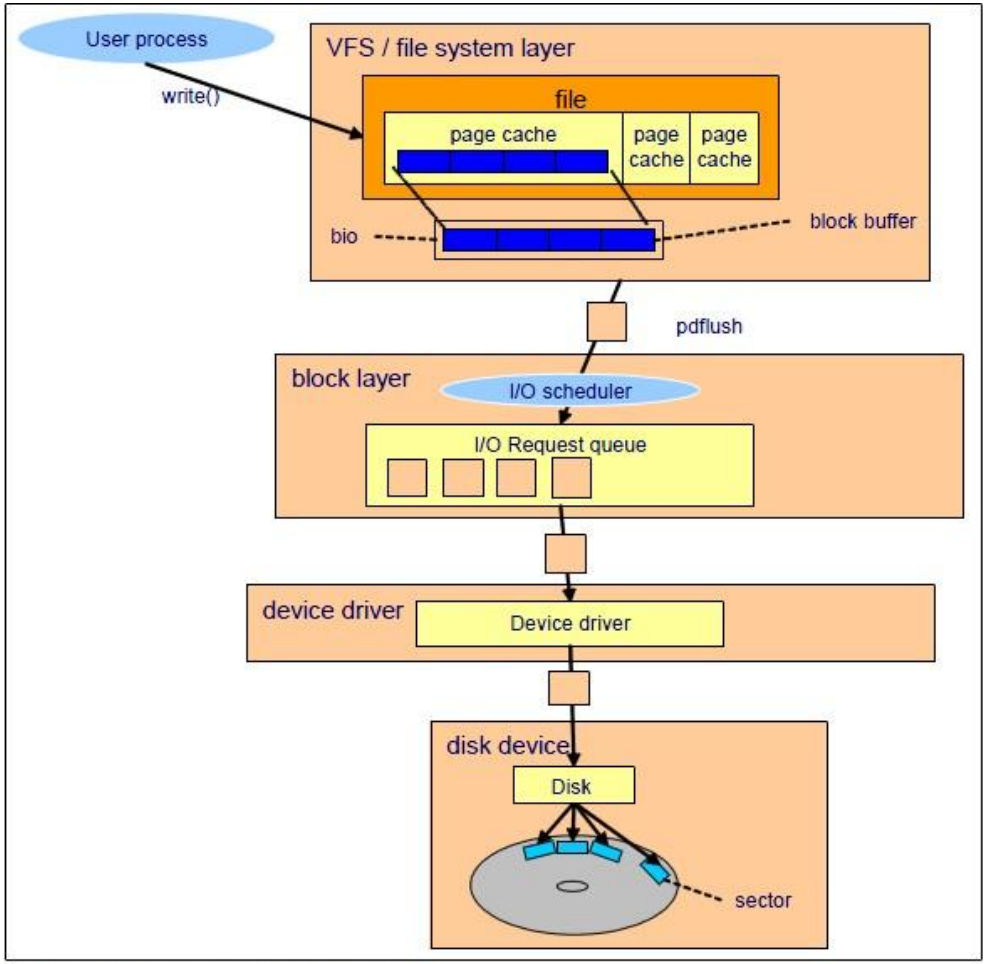
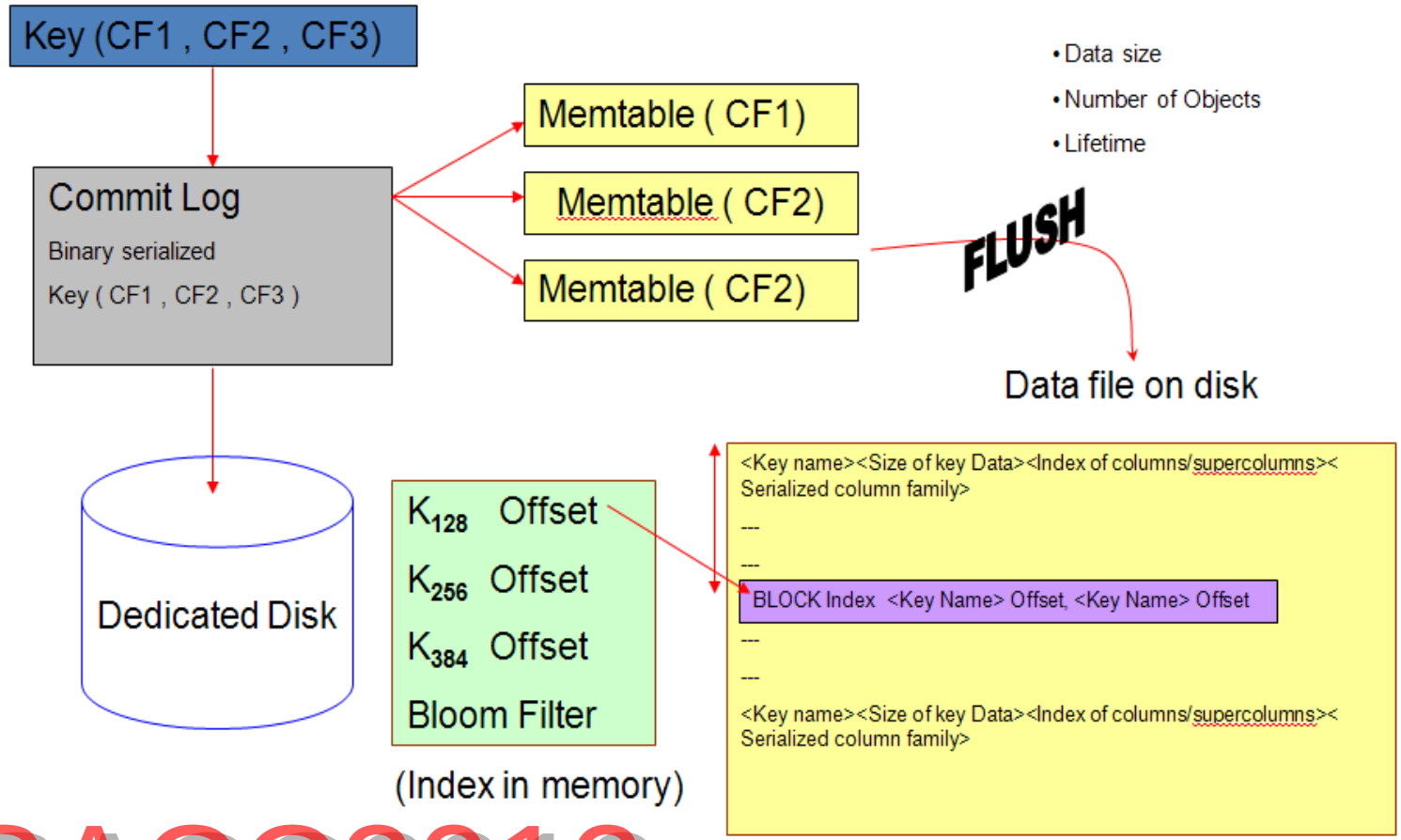


Figure 1-18 I/O subsystem architecture

SACC

数据库缓存--*cassandra* 数据存储过程



有用的性能数字

L1 cache reference	0.5 ns
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	25 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	3,000 ns
Send 2K bytes over 1 Gbps network	20,000 ns
Read 1 MB sequentially from memory	250,000 ns
Round trip within same datacenter	500,000 ns
Disk seek	10,000,000 ns
Read 1 MB sequentially from disk	20,000,000 ns
Send packet CA->Netherlands->CA	150,000,000 ns

存储目标

SACC

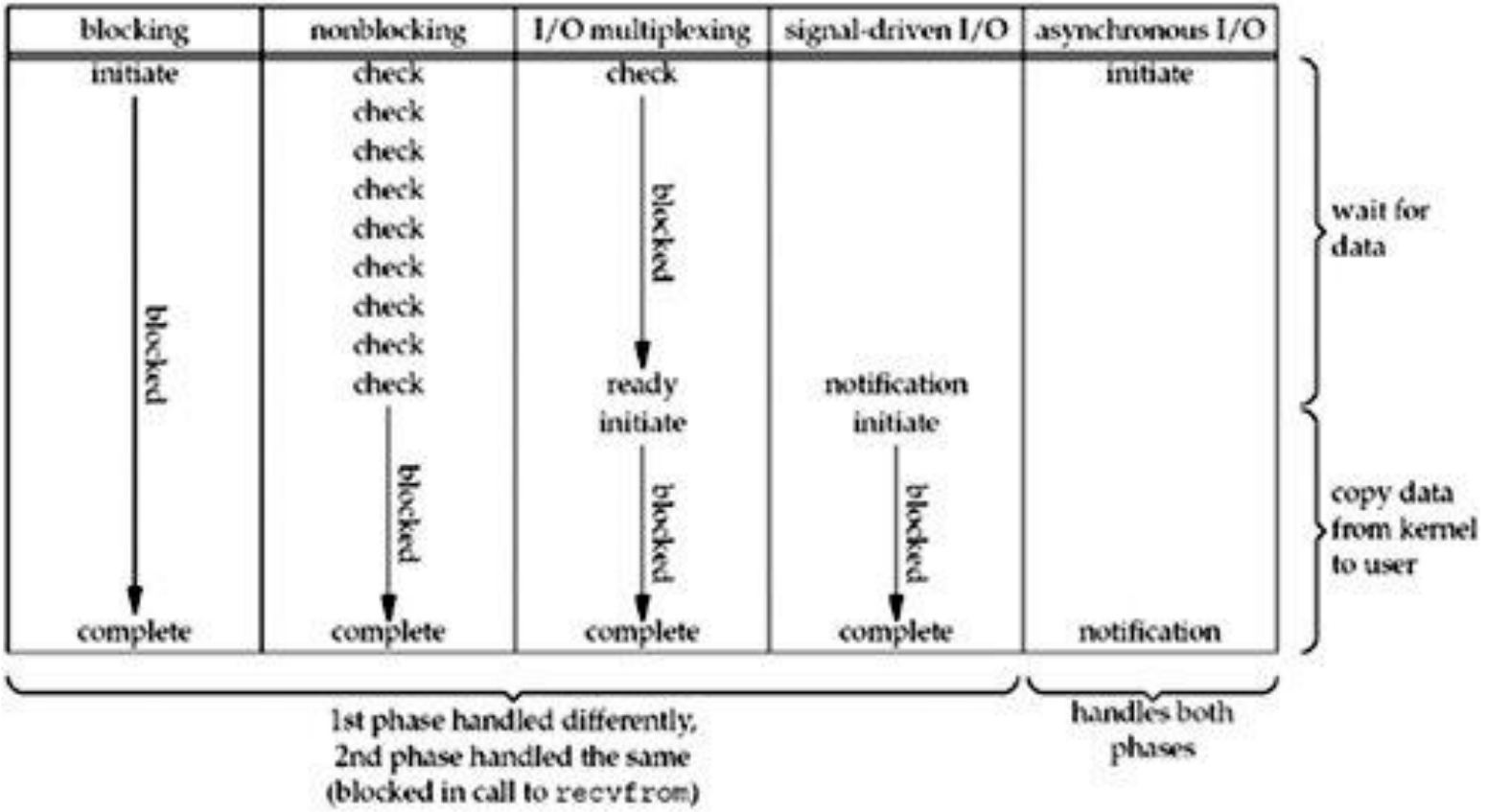
引入：
C10K问题,C500K， C**K

服务器模型与IO模型
s:1 , c:1 , bio; 一个请求一个线程
s:1 , c:n , nio ;多个请求，一个线程分发
seda : Staged Event-Driven Architecture
Select (轮询) 和 epoll （事件驱动 callback）

本质
方法: 查找系统瓶颈 -->Amdahl定律 如IO(sendfile)
目标 :提升性能 -->高吞吐，低延迟
提高资源利用率-->降低成本 -->降低能耗

SACC

IO模型


blocking
nonblocking
I/O multiplexing
signal-driven I/O
asynchronous I/O
initiate
check
check
check
check
check
check
check
check
check
initiate
blocked
blocked
ready
initiate
notification
initiate
blocked
blocked
blocked
complete
complete
complete
complete
notification
wait for
data
copy data
from kernel
to user
1st phase handled differently,
2nd phase handled the same
(blocked in call to recvfrom)
handles both
phases

SACC

分布式基础(存储相关)

分布式存储理论

分布式存储主要存储模型

SACC

分布式存储理论

CAP： Consistency Availability Partition tolerance 只能满足其二

BASE :Basically Available（基本可用）Soft state（柔性状态）

Eventually consistent（最终一致）

Quorum NRW

SACC

分布式存储主要存储模型

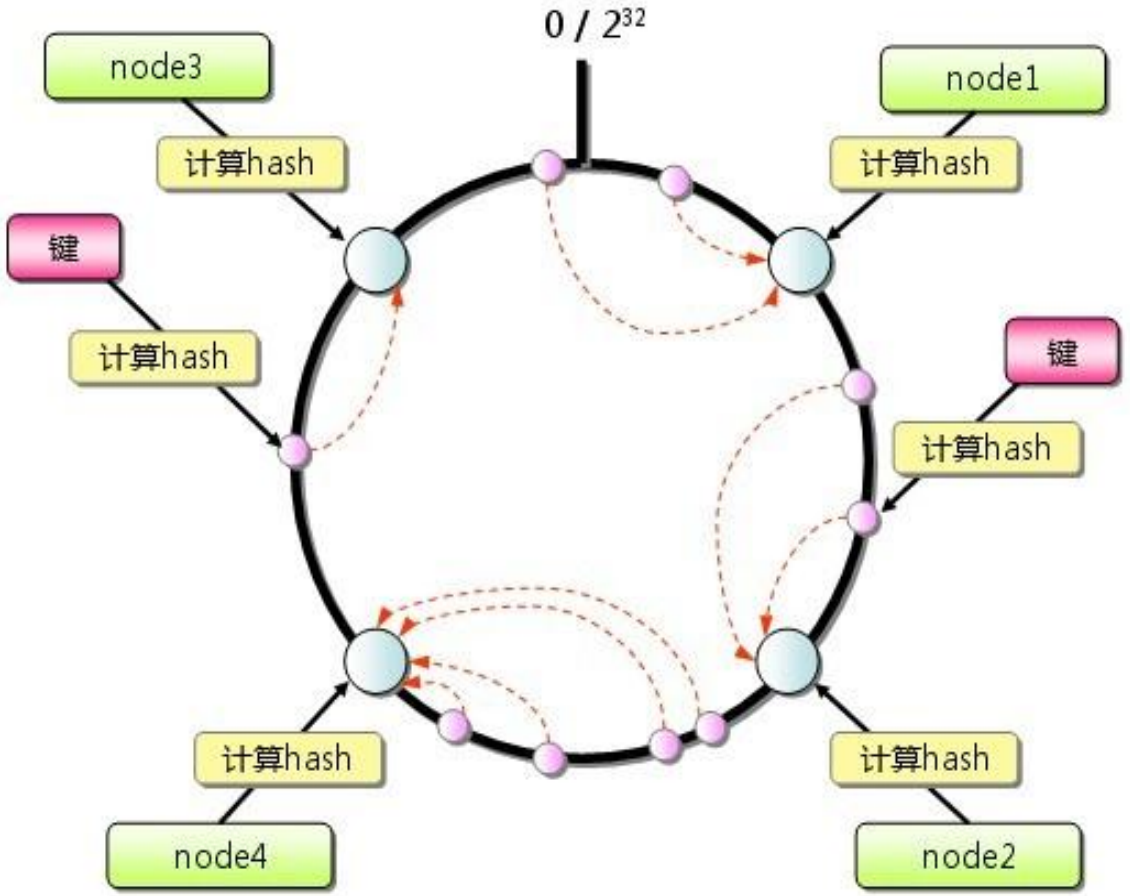
Consistent hash,(去中心化) -->memcahed

B+ tree (实时,随机) -->mongodb

LSM tree, (批量 顺序) -->hbase

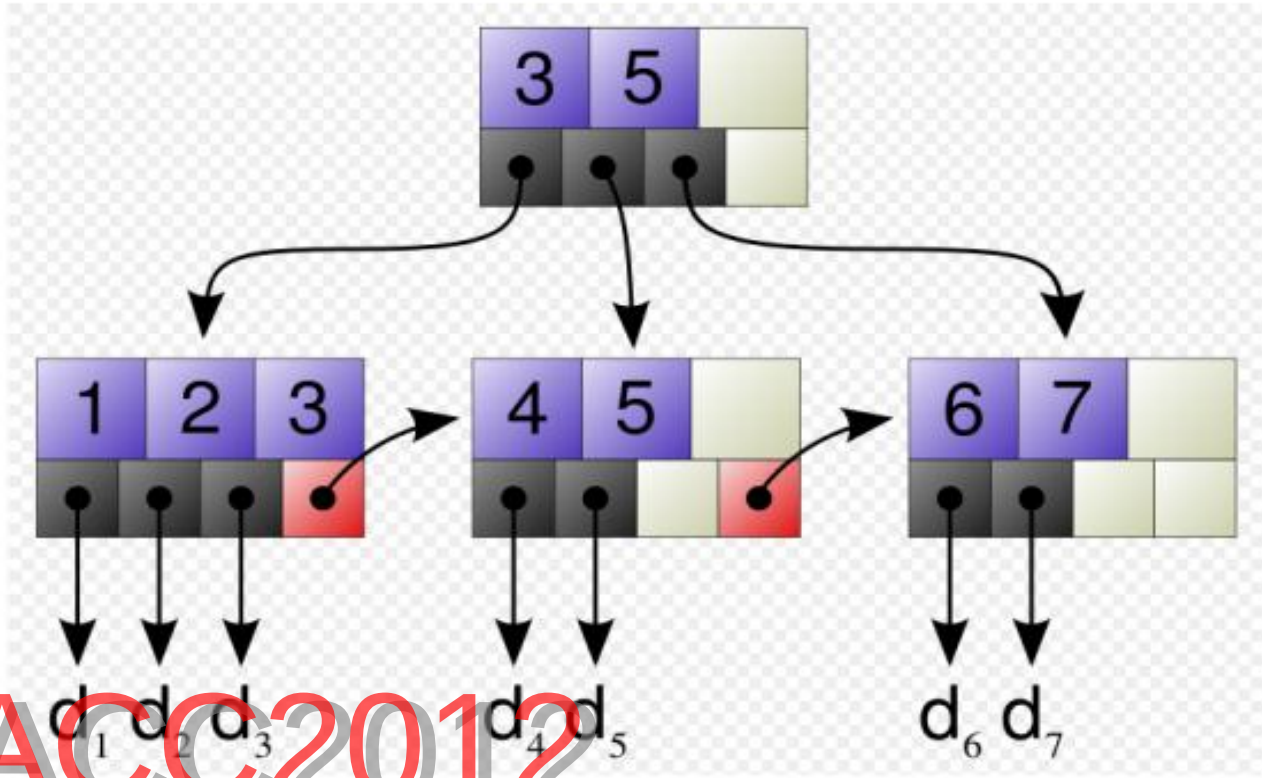
SACC

consistent hash



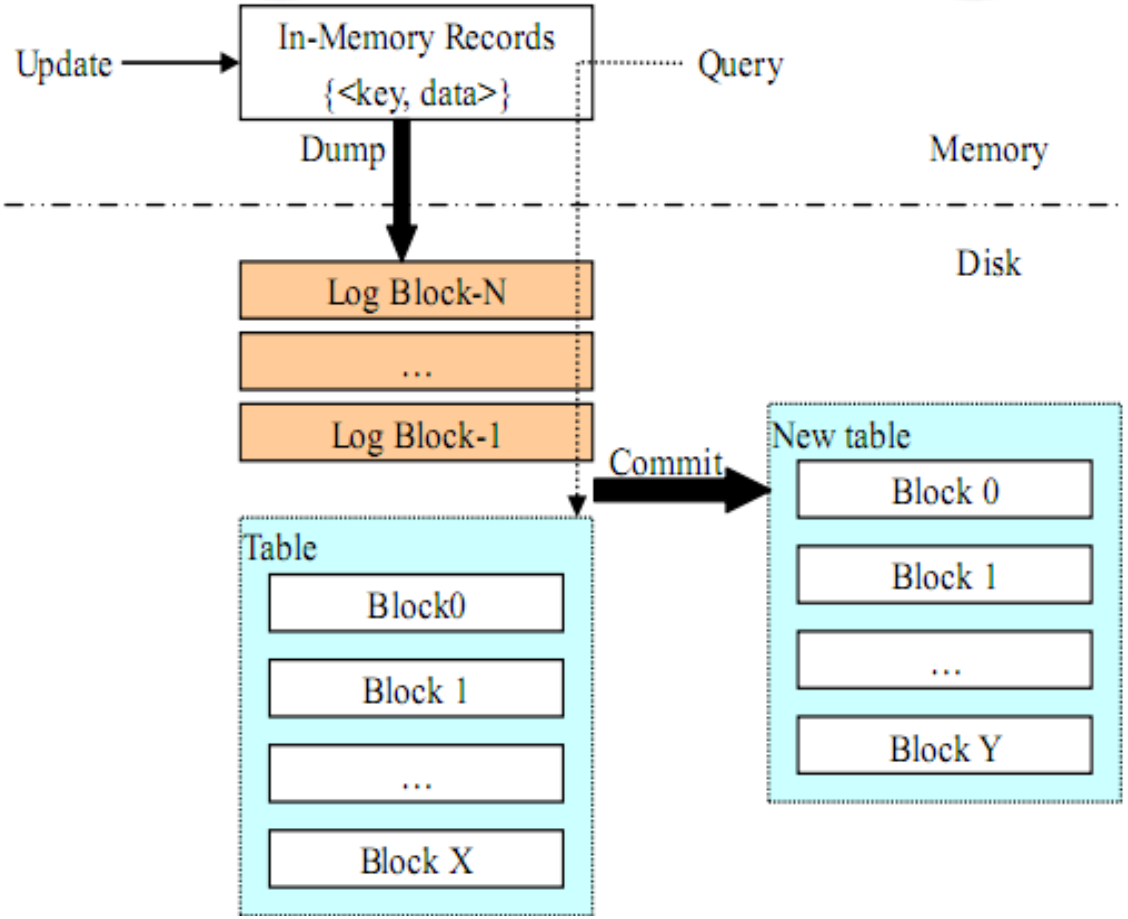
SACC

存储模型：B+ tree



SACC

存储模型：Log-based structure



SACC

如何设计分布式存储架构

分析需求，做好平衡

保持Kiss原则，做到RAS

SACC

分析需求，做好平衡

分析需求

业务场景 核心业务与否 容量大小

数据结构 结构化 半结构化 文件 table , object

访问模式 读写比例，实时读写，顺序读写

做好平衡

CAP选择 , BASE or ACID

选择存储模型 B+ or LSM

资源利用率和管理 吞吐量和延迟 随机与顺序 离线与在线

SACC

原则和目标

原则： kiss

来源 : unix 设计哲学
产品 -->微信摇一摇 苹果新手入门
技术 --->个性化推荐

目标: RAS

RAS: Reliability, Availability, Scalability 高可靠，高可用，高扩展

R:过载控制 ： 管道控制 Qos, (随机早期检测)
A: 容灾 多副本 （同机柜，机房，数据中心）
S 扩容 分片： a 取模; b 一致性hash; c B+ tree 或变种

SACC

58的分布式存储实践

信息系统架构
站内信和统计数(实时)架构
图处存储架构
统计分析平台

实践1：信息系统架构(图)

SACC

SACC

实践1：信息系统架构实践

架构实践:

search engine(index) +Mysql (shard + M/S)+ memcached

分库 : infoid % dbNum

infoid 生成: local times + ip(mac) + pid

扩展: 2的倍数扩展, 备-->主，不用移动数据

改进: 进行通用分布式数据库开发，支持跨库join等

数据量:

- 信息(贴子): ~10亿, 50K qps , memcached 90%hits

- **优点：**

- 简单 成熟 稳定 可靠

SACC

实践2：站内信和统计数(实时)架构

架构实践： mongos + auto sharding（自动分片）

Mongodb 高可用，高性能,线性扩展,无模式，查询支持好

实时统计数服务架构变迁: (mysql + memched) → mysql+ (应用层做缓存)-->cassandra-->redis--> mongodb(线性扩展)

升级为通用服务:appid+appinforid == _id （key最小）

分片: 站内信（用户ID） 统计数(信息ID)

站内信: range-->kv ， sql兼容

数据量： 站内信 : ~2亿 统计数: ~10亿, 20kqps

优点: 简单 高扩展 高吞吐

实践3：图片存储示例


58.com 团购
请输入商品、商家名称
搜索
稻香村 中秋礼品 大闸蟹 床上四件套 秋装
本地特色月饼
思乡送礼啦！
首页 餐饮美食 休闲娱乐 生活服务 旅游酒店 网上购物 品牌特卖
登录 注册
类别： 全部 中餐 (512) 酒店 (384) 火锅 (241) 休闲餐厅 (101) 综合游乐场 (99) KTV (88) 西餐 (83) 自助烤肉 (77) 西点屋 (65) 韩式烤肉 (62)
按摩推拿 (59) 传统烧烤 (51) 健身卡 (50) 足疗 (45) 景点门票 (42) 婚纱摄影 (37) 美发 (36) 桌球 (36) 温泉 (36) 电影票 (34) 海鲜套餐 (34) 更多
网购： 全部 服装服饰 (169) 家居日用 (82) 汽车户外 (58) 箱包饰品 (51) 休闲食品 (50) 鞋靴 (45) 母婴用品 (18) 数码家电 (12)
商圈： 全部 朝阳 (847) 海淀 (578) 丰台 (371) 西城 (242) 东城 (207) 昌平 (147) 崇文 (111) 通州 (84) 宣武 (79) 大兴 (65) 石景山 (57) 顺义 (51) 更多
今日推荐
关注最多
销售最多
猜您喜欢
三只小老虎。
商圈：安贞
3.78折： 中国木偶剧院大型现代童话剧《三只小老虎》真人表演门票
¥68
去看看
原价：¥180 节省：¥112 150人已购买
超凡蜘蛛侠
17.5影城
四店通用
商圈：朝阳 小西天 皇君庙 万寿路
2.88折： 4店通用！仅23元！抢购『17.5影城』2D单人观影票1张
¥23
去看看
原价：¥80 节省：¥57 401人已购买
免运费
1.99折： 【全国包邮】仅119元抢全棉斜纹印花四件套一套！35款任选
¥119
去看看
原价：¥599 节省：¥480 380人已购买

SACC

实践3：图片存储架构实践

架构实践：cdn (Squid 网络延迟) + (lvs)+ Ngix (代理,实时生成缩略图) +httpServer(接入层， webdav,sso) + simple GFS(master-slave)

学习: Facebook开源服务器、数据中心,开源存储方案

计算: GraphicsMagick ; opencl

扩展: rest的URI层次扩展,文件名携带所有的信息

备份: 三份， 主 + 实时备 +（延时备份 不同机房）

分层: ram -->ssd-->sas (数据访问热度) -->lsm

数据量：total 50t ，20亿record , 100G/add ; 20:1 (r:w);

优点 性能高 成本低 易扩展

SACC

实践3：附facebook数据中心电力布局

SACC

实践4：统计分析平台

架构实践: Hadoop + zookeeper +hbase+redis+mongodb+...

特点 :友好的用户界面,支持多种数据源;

只写部分业务代码,即可以运行和调度

场景:如用户,击行为分析 （按时间,区域统计信息点击数）

HBASE： 高可靠性,高性能, 面向列,可伸缩的分布式存储系统

强一致性, 海量数据

Redis : 支持丰富数据结构 如list (数据挖掘)

HDFS : 读取日志等大文件

改进： 自主开发Drm云平台

数据量: ~100t

优点：简单易用 满足多样性需求

微博：http://weibo.com/zhuozhe
@浊者
邮箱：xuzh@58.com

Thanks !