

系统设计 | 场景 | 海量数据 | 智力

一些未归档的点

- 数据库与缓存双写一致性
 - <https://www.cnblogs.com/crazymakercircle/p/14853622.html#autoid-h2-15-11-0>
- Redis分布式锁
 - <https://www.cnblogs.com/crazymakercircle/p/14731826.html#autoid-h2-14-8-0>
- Redis集群
 - <https://www.cnblogs.com/crazymakercircle/p/14698576.html#autoid-h3-15-0-0>
- 分布式
 - CAP、BASE
 - 一致性协议：Paxos、2PC、Raft、ZAB
 - 分布式消息队列、分布式缓存
 - 分库、分表、uuid
 - 分布式事务：
 - 分布式事务他是作用于多个节点上的，因为现在很多都是微服务，不同的服务部署在不同机器上，需要保证某个操作的原子性的时候就需要分布式事务；
 - 比如你下一个订单，你需要经过订单系统生成订单，还有库存系统减少库存，还有你的个人账户要扣钱；
 - 这三个动作要么都成功，要么都失败
 - 这就是分布式事务所要保证的东西
 - XA、TCC、SAGA
 - <https://www.cnblogs.com/crazymakercircle/p/13917517.html#autoid-h3-29-0-0>
 - <https://www.cnblogs.com/crazymakercircle/p/14334422.html#autoid-h2-7-1-1>
 - Raft: <https://www.cnblogs.com/crazymakercircle/p/14343154.html#autoid-h2-4-8-0>
 - 分布式系统遇到的十个问题 悟空聊架构
- 限流算法（计数器、漏桶、令牌桶）
 - <https://www.cnblogs.com/crazymakercircle/p/15187184.html>
- uuid(唯一id)生成（雪花算法）
 - <https://cloud.tencent.com/developer/article/1530850>
- CDN（内容分发网络）
 - <https://www.cnblogs.com/crazymakercircle/p/14978513.html#autoid-h2-1-0-0>
- DNS超详解
 - <https://www.cnblogs.com/crazymakercircle/p/14976612.html#autoid-h2-17-0-0>

系统设计

参考文章

- [架构设计面试题（史上最全、持续更新、吐血推荐）](#)

秒杀系统

- [万字长文，解密秒杀架构！（建议收藏）](#)
- CDN
- 缓存+数据库
- 读写分离
- 答题校验模式
- 业务隔离、系统隔离、数据隔离

微信摇一摇设计（两个子问题：同一秒的用户配对到一起 + 保证用户的唯一分配，b配给a的时候保证不配给c）

- 按秒划分
- 并发安全：锁、原子操作（CAS底层实现）

如何查找附近的人（两个子问题：确认互相匹配 + 推荐到离自己距离多远的人）

- 哈希表
- [GEO哈希](#)

双十一流量尖峰

- Redis保证MySQL安全
- 缓存预热、服务熔断、服务降级
- 主从读写分离
- RocketMQ队列下单
- 防止恶意请求和爬虫：答题机制

抢红包算法

- $2 * \text{剩余金额} / \text{剩余红包数}$
- [四种方法讲解](#)

给微信发红包功能设计[测试](#)用例？

- 文章：<http://www.51testing.com/html/34/n-4478734.html>

设计点赞功能（feed流）

- [feed流文章](#)（好文好文）

让你去设计一个可靠的通信协议你去怎么设计

- 应用层的数据的序列化
- 定义好包的格式
- 还有数据的加解密
- 增加 CRC 做包的完整性校验
- 协议的版本控制
- 超时机制等等

设计一个 缓存系统，10万条<key,val> 存到五台机器上如何设计？

- 分布式缓存、Redis
- rehash如何解决：

todo：

- 分布式id、短链、feed流、微博、微信红包、日志监控、直播弹幕
- 考察：需求分析、存储设计、服务交互、注重高性能、高可用、高扩展

场景题

现在内存大小16G，malloc(10G)，会发生什么，如果现在再加一条语句给这些内存赋初值，又会发生什么

总结

至此， 验证完成了。简单总结下：

- 在 32 位操作系统，因为进程最大只能申请 3 GB 大小的虚拟内存，所以直接申请 8G 内存，会申请失败。
- 在 64位 操作系统，因为进程最大只能申请 128 TB 大小的虚拟内存，即使物理内存只有 4GB，申请 8G 内存也是没问题，因为申请的内存是虚拟内存。如果这块虚拟内存被访问了，要看系统有没有 Swap 分区：
 - 如果没有 Swap 分区，因为物理空间不够，进程会被操作系统杀掉，原因是 OOM（内存溢出）；
 - 如果有 Swap 分区，即使物理内存只有 4GB，程序也能正常使用 8GB 的内存，进程可以正常运行；

- [小林讲解](#)

服务器A向B发包 如果遇到网络阻塞都会发生什么

- 阻塞、超时需要重传
 - 没发出去、中途丢掉、缓冲区满丢掉、收到在网卡但没被拉上去
- 重传需要保证幂等
 - 幂等指的是相同请求（identical request）执行一次或者多次所带来的副作用（side-effects）是一样的。
 - 加一个syn字段，对于重发的，syn相等，丢弃

牛客上代码运行的过程，换句话说点击QQ到页面打开的过程，从操作系统的层面思考

- 操作系统检测类型是否是可执行文件，Windows的可执行文件是PE
- 创建进程，并且将可执行文件映射到该进程
- 为该进程设置CPU上下文环境
- 将代码和数据从磁盘读入内存
- 运行过程中发生缺页异常则重复4
- 执行相关系统调用函数，在显示器上显示图标

为了保证网络接口的安全性，需要做哪些措施

- 数据要加密
- 内容防止篡改：hash md5、sign

- 身份验证：Token、数字签名
- 防止重复提交，幂等：
- [文章1](#)
- [文章2](#)

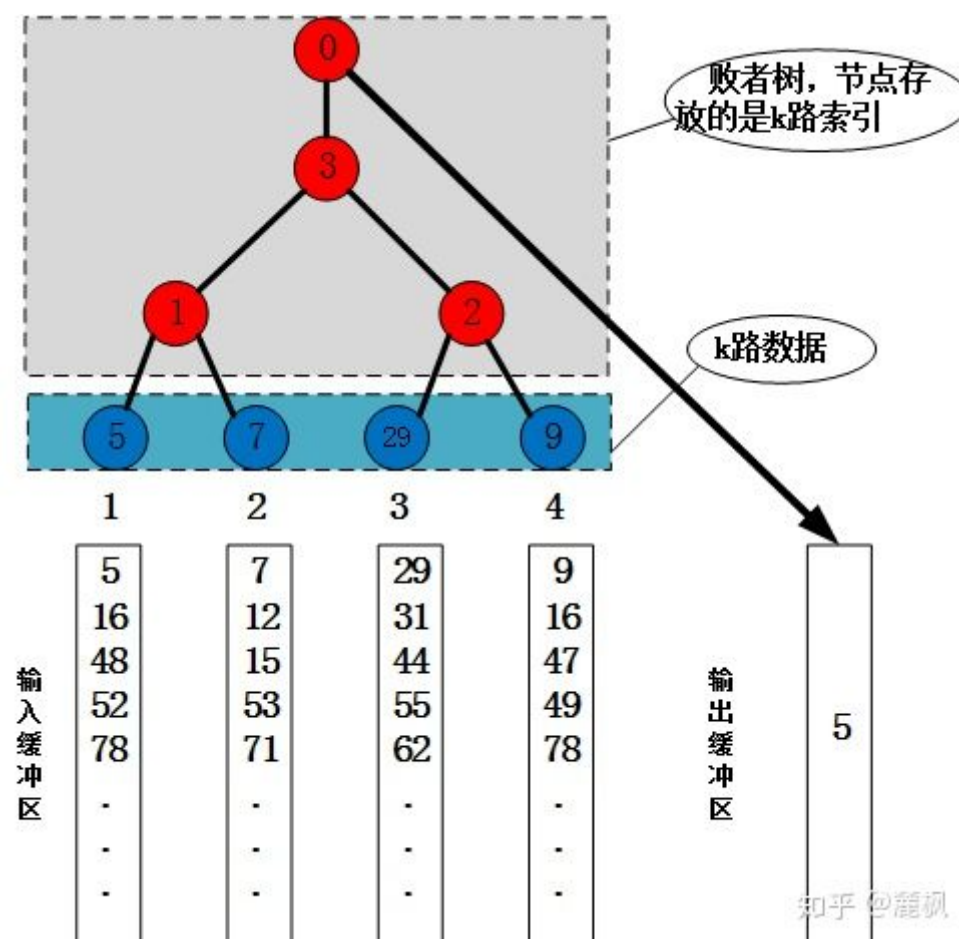
海量数据

- 参考文章：
 - [阿秀总结](#)
 - [10 道 BAT 大厂海量数据面试题（附题解+方法总结）](#)
 - [十个面试常问海量数据处理方法总结\(建议收藏\)](#)
 - [这些海量数据类题，面试出现率极高，请收藏！](#)
- Conception：
 - a. 哈希取余
 - b. 分治
 - c. hashmap
 - d. Trie树
 - e. 小顶堆大顶堆（前100大的用**小顶堆**筛）
 - f. Bitmap
 - g. 布隆过滤器
 - h. TopK（快速选择、堆）
 - i. 数据库索引、倒排索引（Inverted index）
- Question：
 - 8G内存，50G的文件，如何去重
 - 对文件进行MD5哈希，用一个<哈希值，文件名>进行去重
 - 给定 a、b 两个文件，各存放 50 亿个 URL，每个 URL 各占 64B，内存限制是 4G。请找出 a、b 两个文件共同的 URL
 - 有一个 1GB 大小的文件，文件里每一行是一个词，每个词的大小不超过 16B，内存大小限制是 1MB，要求返回频数最高的 100 个词(Top 100)
 - 现有海量日志数据保存在一个超大文件中，该文件无法直接读入内存，要求从中提取某天访问百度次数最多的那个 IP
 - 在 2.5 亿个整数中找出不重复的整数
 - 给定 40 亿个不重复的没排过序的 unsigned int 型整数，然后再给定一个数，如何快速判断这个数是否在这 40 亿个整数当中
 - 搜索引擎会通过日志文件把用户每次检索使用的所有查询串都记录下来，每个查询串的长度不超过 255 字节。假设目前有 1000w 个记录（这些查询串的重复度比较高，虽然总数是 1000w，但如果除去重复后，则不超过 300w 个）。请统计最热门的 10 个查询串，要求使用的内存不能超过 1G。（一个查询串的重复度越高，说明查询它的用户越多，也就越热门。）
 - 已知某个文件内包含一些电话号码，每个号码为 8 位数字，统计不同号码的个数
 - 从 5 亿个数中找出中位数。数据排序后，位置在最中间的数就是中位数。当样本数为奇数时，中位数为第 $(N+1)/2$ 个数；当样本数为偶数时，中位数为第 $N/2$ 个数与第 $1+N/2$ 个数的均值。

- 从高位往地位进行遍历，二分查找
- 有 10 个文件，每个文件大小为 1G，每个文件的每一行存放的都是用户的 query，每个文件的 query 都可能重复。要求按照 query 的频度排序。
- 有 20 个数组，每个数组有 500 个元素，并且有序排列。如何在这 20*500 个数中找出前 500 的数

外部排序

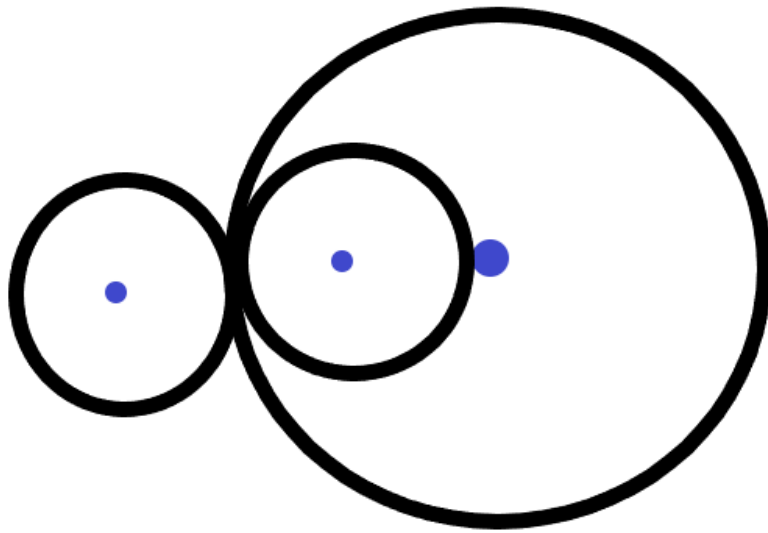
- 参考文章
 - [编程之法：外排序](#)
- Conception:
 - 多路归并、胜者树、败者树



智力题

- 100瓶水，一瓶有毒，有一种试纸，不过需要一个小时才能出结果，问最少需要几片试纸才能在一小时内找到有毒的那一瓶？
 - 7张，7张能表示127个数字，100的二进制是110 0100
 - 布隆过滤器：查一个数所对应的所有1位是否为1（布隆过滤器有误差）
- 有两位盲人，他们都各自买了两对黑袜和两对白袜，八对袜子的布质、大小完全相同，而每对袜子都有一张商标纸连着。两位盲人不小心将八对袜子混在一起。他们每人怎样才能取回黑袜和白袜各两对呢？
 - 随机拿，拿完每双拆开，交换
 - 黑色比白色吸热，黑色干的快
 - [讲解](#)
- 你有四个装药丸的罐子，每个药丸都有一定的重量，被污染的药丸是没被污染的药丸的重量+1。只称量一次，如何判断哪个罐子的药被污染了？
 - 1、2、3、4四个罐子分别取1、2、3、4个药丸，称总重量
 - 罐子编号 = 总重量 - 实际1、2、3、4药丸的重量

4. 有十个机器，九个是生产的金币是5g，只有一个生产的是4g，给你一个称，问你一次怎么找出来那个生产4g金币的机器？
- 第i个机器拿i个金币，混到一块称总重量，少多少g就是第几台机器
5. 两个圆环，半径分别是1和2，小圆在大圆内部绕大圆圆周一周，问小圆自身转了几周？如果在大圆的外部，小圆自身转几周呢？



- 2周
6. 有两根不均匀分布的香，香烧完的时间是一个小时，你能用什么方法来确定一段15分钟的时间？
- 一根点一端，一根点两端
 - 点两端的香烧完的时候是半个小时
 - 点了一端的一根开始两端都点，烧完时候就是15分钟
7. 一群人开舞会，每人头上都戴着一顶帽子。帽子只有黑白两种，黑的至少有一顶。每个人都能看到其它人帽子的颜色，却看不到自己的。主持人先让大家看看别人头上戴的是什么帽子，然后关灯，如果有人认为自己戴的是黑帽子，就打自己一个耳光。第一次关灯，没有声音。于是再开灯，大家再看一遍，关灯时仍然鸦雀无声。一直到第三次关灯，才有劈劈啪啪打耳光的声音响起。问有多少人戴着黑帽子？
- 黑的至少有一顶
 - 第一次没声音，说明 >1 顶，带黑帽子的人看见还有1顶黑帽子才不会打
 - 第二次没声音，说明 >2 顶，根据第一次已经知道最少2顶，带黑帽子的人看见还有2顶黑帽子才不会打
 - 第三次有声音说明为3顶，带黑帽子的人根据第二轮 >2 没人打可以推测自己是黑帽子
 - 结论：第几次关灯就有几顶黑帽子
8. 有一个水龙头，一个5L水杯，一个7L水杯，如何精准地获得6L水？
- 7升装满，倒进5升里，剩2升
 - 2升转移进5升里
 - 7升装满，倒进5升里，剩4升，5升倒
 - 4升转移进5升里
 - 7升装满，倒进5升里，剩6升
9. 一个圆形棋盘，双方分别下子，采取什么策略能赢

