

# 系统设计

## Distributed System Design 2

### (九章网站下载最新课件)

本节主讲人: 北丐

版权声明: 九章课程不允许录像, 否则将追究法律责任, 赔偿损失



扫描二维码关注微信/微博  
获取最新面试题及权威解答

微信: [ninechapter](#)

微博: <http://www.weibo.com/ninechapter>

知乎: <http://zhuankan.zhihu.com/jiuzhang>

官网: <http://www.jiuzhang.com>

- Design a Bigtable
  - Google, Facebook, Amazon, Alibaba
  - NoSQL database 设计框架和原理
  - SStable 读和写
  - Bloom Filter

# Interviewer: What is bigtable?

$\aleph_0$  脸茫然  
 $\left\{ \text{脸} \right\}_{n=0}^{\infty}$

# What is bigtable?

| NoSQL DataBase | Company                               |
|----------------|---------------------------------------|
| Bigtable       | Google                                |
| Hbase          | Yahoo(Alitaba)Open Source of Bigtable |
| Cassandra      | Facebook                              |
| DynamoDB       | Amazon                                |

Comparison of different No SQL database:

<https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models>

# 为什么我们要讲bigtable 的实现？

1. Google面试题
2. 解决相类似系统设计题,比如:Look up service
3. 追问NoSQL How to scale的原理



# 文件系统 vs 数据库系统

# 什么是文件系统？

操作：

输入：/home/jinyong/character\_name.txt

输出：文件内容

# 如果有下面需求 找到“令狐冲”的“颜值”

- 1、打开文件
- 2、For循环扫描文件的内容  
然后找令狐冲的颜值

```
{  
{ '姓名': '令狐冲', '颜值': 5, '身高':  
'160cm'},  
{ '姓名': '郭靖', '颜值': 9, '身高': '180cm'},  
{ '姓名': '东邪', '颜值': 7, '身高': '170cm'},  
}
```

/home/jinyong/character\_name.txt



# 文件系统缺点？

文件系统提供一些简单的读写文件操作

实际查询当中有复杂的查询需求：

比如：查询令狐冲颜值

查询颜值小于5的

查询所有人的身高平均值

所以我们需要一个更复杂的系统建立在文件系统之上

# 数据库系统

- 1、建立在文件系统之上
- 2、负责组织把一些数据存到文件系统
- 3、对外的接口比较方便操作数据

# 什么是数据库系统

操作

输入：key(令狐冲+颜值)

输出：value (5)

查询令狐冲颜值

| 姓名  | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |
| 郭靖  | 9  | 180 |
| 东邪  | 7  | 170 |

# 设计数据库系统

# Scenario 需求

查询: key (令狐冲 + 颜值)

返回: value (5)

因为后端系统通常给web server使用, Scenario比较单一

# Storage

## 数据库怎么存储 以表的形式？Yes/No

| 姓名  | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |
| 郭靖  | 9  | 180 |
| 东邪  | 7  | 170 |

# 数据最终都会存到文件里面

```
{  
  {'姓名': 'linghuchong', '颜值': 5, '身高':  
    '160cm'},  
  {'姓名': 'guojing', '颜值': 9, '身高':  
    '180cm'},  
  {'姓名': 'dongxie', '颜值': 7, '身高':  
    '170cm'},  
}
```

# 从文件系统基础上思考 搭建数据库系统



# 在文件里面 怎么更好支持查询操作？

```
{  
  {'姓名': 'linghuchong', '颜值': 5, '身高':  
    '160cm'},  
  {'姓名': 'guojing', '颜值': 9, '身高':  
    '180cm'},  
  {'姓名': 'dongxie', '颜值': 7, '身高':  
    '170cm'},  
}
```

# 读取文件到内存里面 内存里面查询

有什么问题？

# 先对数据进行排序？

不需要一个一个在文件里面查询

只需要二分扫描

那么文件存储在硬盘里面，怎么在硬盘里面进行二分呢？

# 硬盘二分查询

画图解释硬盘二分

Read More:

<http://stackoverflow.com/questions/736556/binary-search-in-a-sorted-memory-mapped-file-in-java>

# 查询解决了， 有一天令狐冲整容了怎么办？

修改令狐冲颜值，从5变到6

1. 直接在文件里面修改
2. 读取整个文件，修改好了，重新写入覆盖原来文件
3. 不修改，直接append操作加在文件最后面

### 1. 直接在文件里面修改

很难做到直接修改内容, 如果原来是4个字节, 现在修改成8个字节, 那么之后的内容都需要移动位置。

### 2. 读取整个文件, 修改好了, 重新写入覆盖原来文件

非常耗费时间, 每次要读出写入 其他多余不变的内容

### 3. 直接append操作加在文件最后面

好处: 特别快

# Bigtable为了写优化 选择了直接Append

坏处：读取数据怎么办，文件没有顺序，？

# 读取数据怎么办？

把所有关于令狐冲的颜值记录都读出来



# 怎么解决文件没有顺序 不能二分查询的问题？

过一段时间统一把文件有序整理

有木有一个办法  
既可以读的时候二分查询  
又可以写的时候写在最后append操  
作？

分块有序

1. 每一块都是内部有序
2. 写的时候只有最后一块是无序的

# 块会越写越多，会有很多重复

(令狐冲经常做整形手术)

重复非常多

每次查询所有的块非常消耗时间

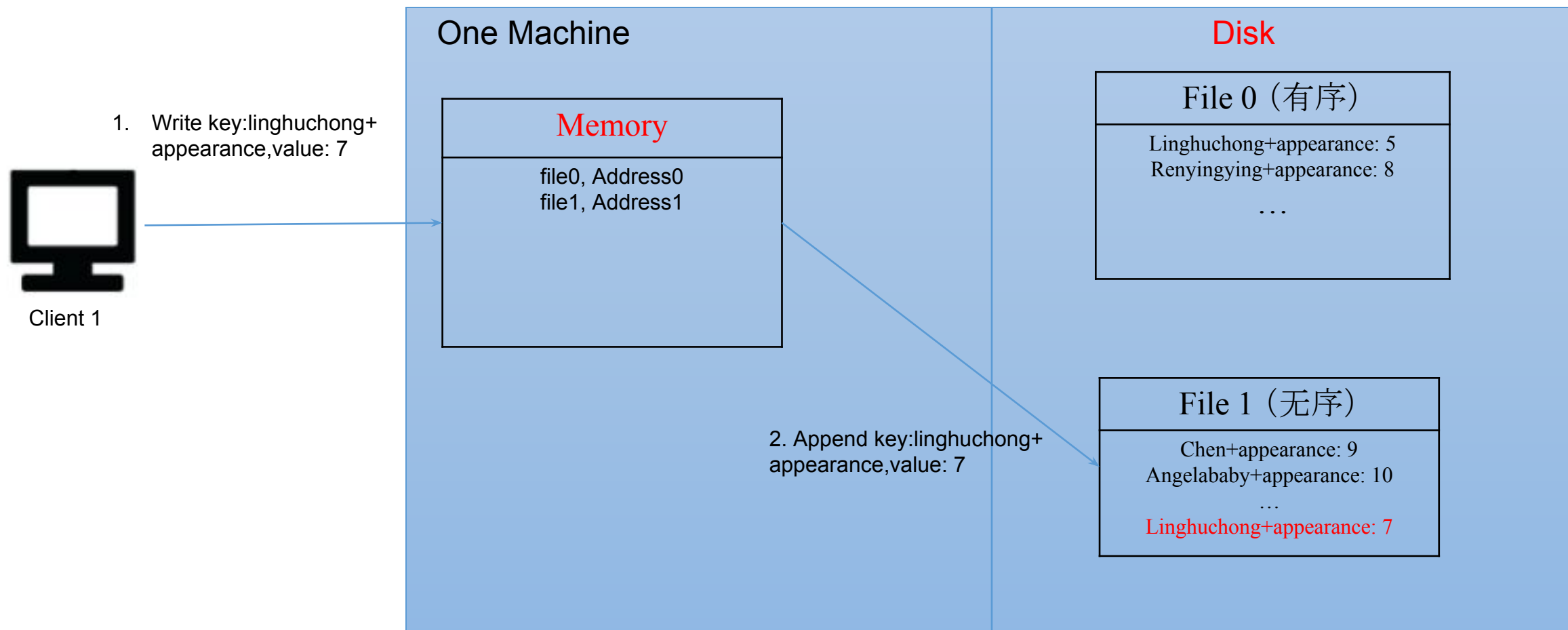
定期K路归并

<http://www.lintcode.com/en/problem/merge-k-sorted-arrays/>

# 完整系统读/写过程

One Work Solution

# 写入过程



# 怎么把最后一个File 从无序变成有序？

1. 读入到内存快速排序
2. 硬盘外部排序
3. 可不可以一开始就存在内存里面？

1. 读入到内存快速排序。

所有数据1次硬盘写入, 1次硬盘统一读取+内存排序+1次硬盘统一写入

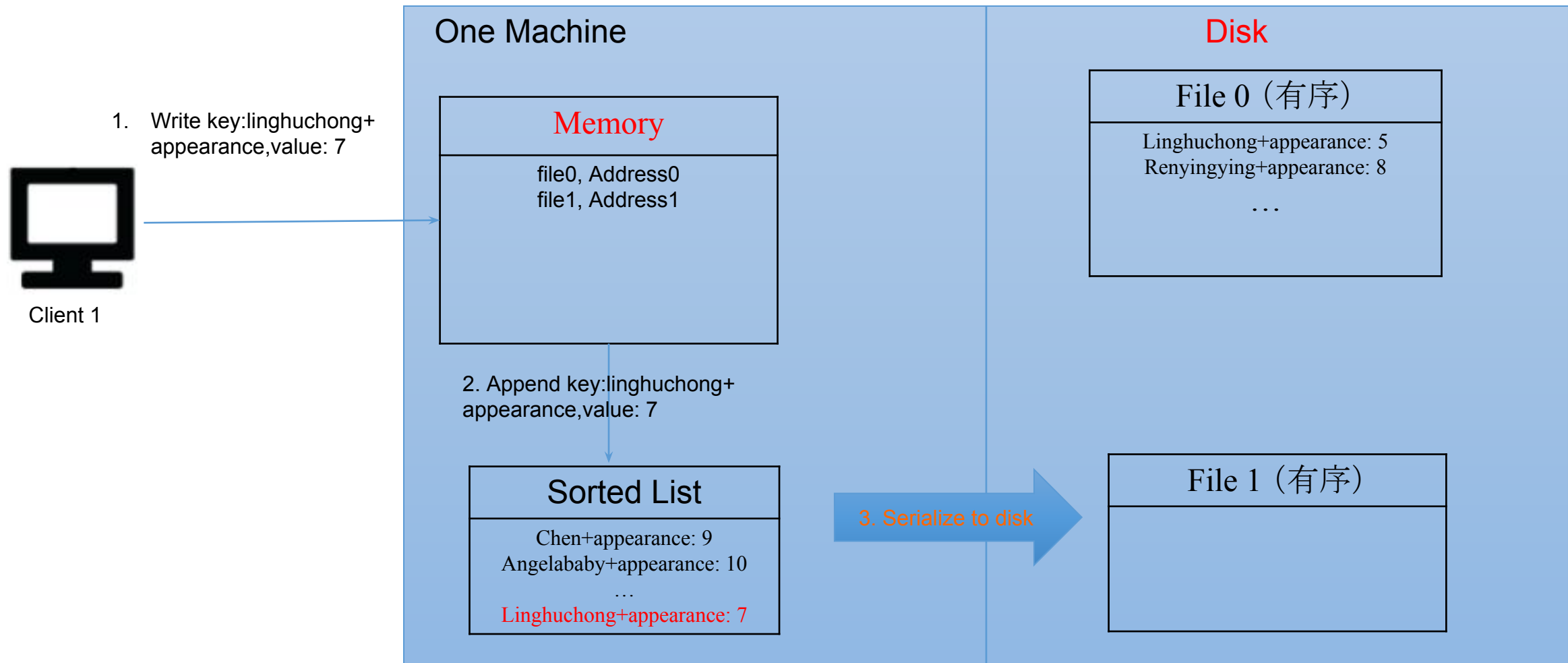
2. 硬盘外部排序

有必要么？

3. 可不可以一开始就存在内存里面？

内存排序+1次硬盘写入





# Serialization

<http://www.lintcode.com/en/problem/binary-tree-serialization/>

Interviewer: 机器挂了，内存没了？

# Write Ahead Log (WAL)

那写log岂不是又要写硬盘

# WAL 非常方便,不像 重要数据需要整理

内存排序+1次硬盘统一写入+1次硬盘写Log

Link:

<http://www.larsgeorge.com/2010/01/hbase-architecture-101-write-ahead-log.html>

# 读出过程



Client 1

1. Read key:linghuchong + appearance,

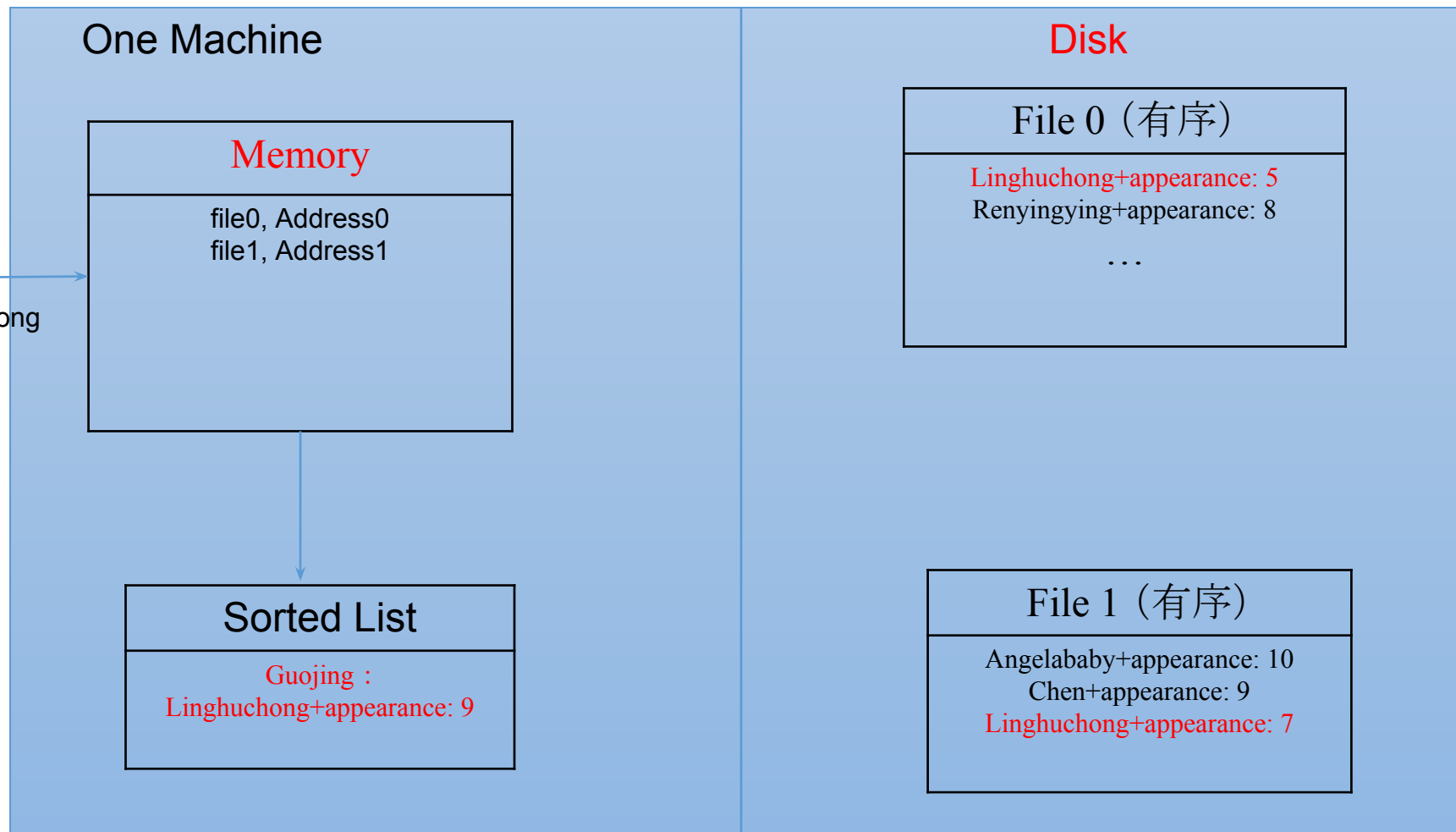
Question:

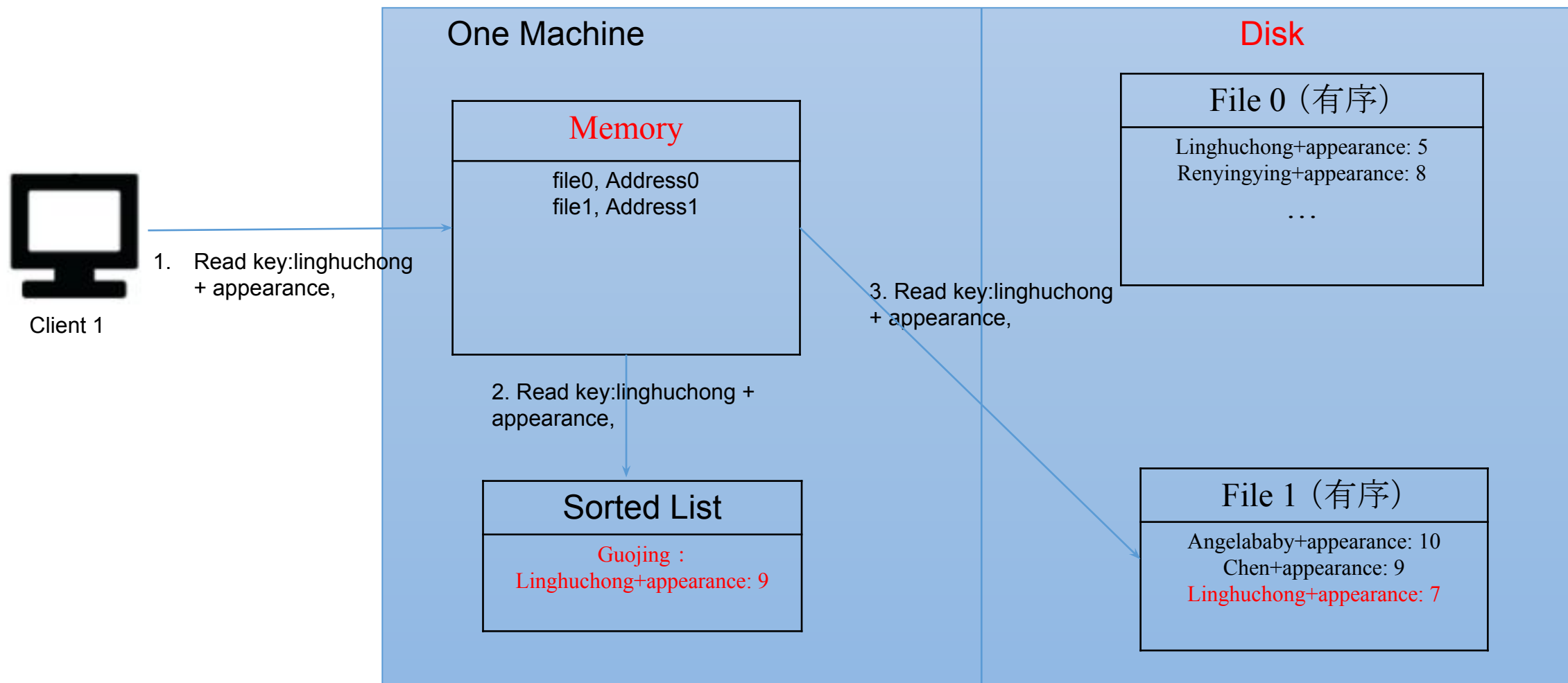
Sorted List

File 0

File 1

寻找【真】令狐冲？







# 一个File里面怎么查询令狐冲？

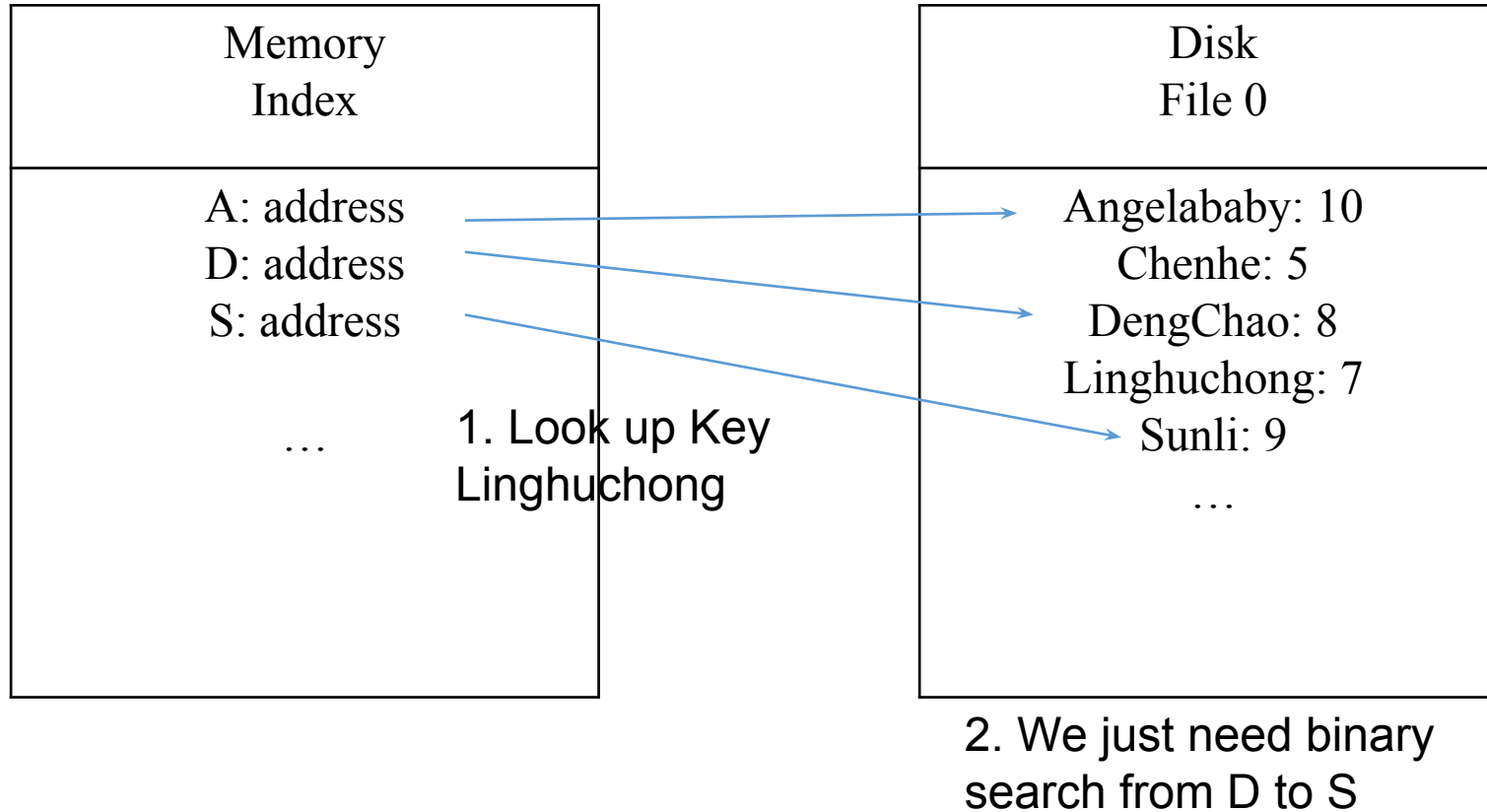
1. 读出来for循环
2. 硬盘二分
3. 有木有更好的方法？

# 建立Index

目的：加快查询

怎么建？

# One easy way to build index



## Key

- 把一些Key放入内存作为Index
- Index有效减少磁盘读写次数

# Intersection of Two Arrays ii Follow Up

<http://www.lintcode.com/en/problem/intersection-of-two-arrays-ii/>

这道题的challenge题目

Read More:

B tree index: <http://bit.ly/2bTwhZC>

休息5分钟

# 有木有更好的方法检查一个key在不在一个File里面？

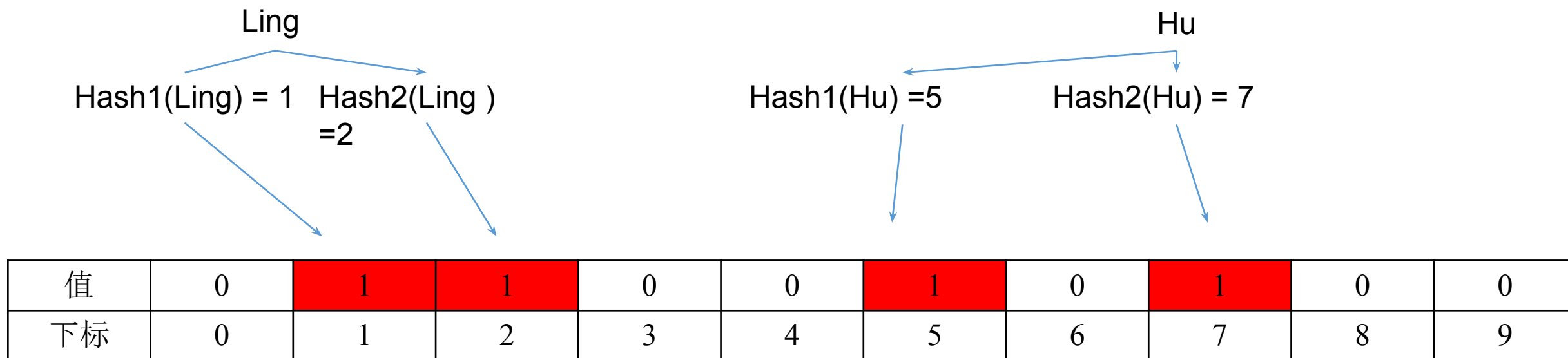
为什么要做如此多的读优化？

因为在写的时候做了Append优化，才会想办法加快读的速度。

# BloomFilter

# Interview: How to look up in bloom filter?

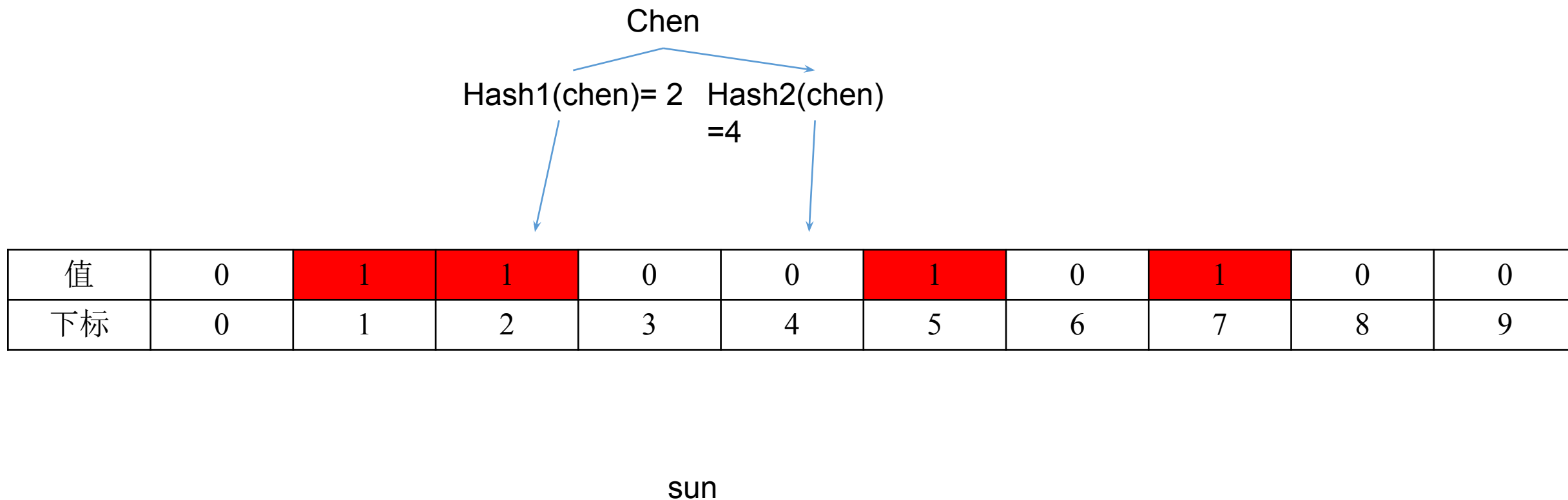




Chen LI

## Interview: How to find in bloom filter?

- 如何检查“chen” 在bloom filter 里面？



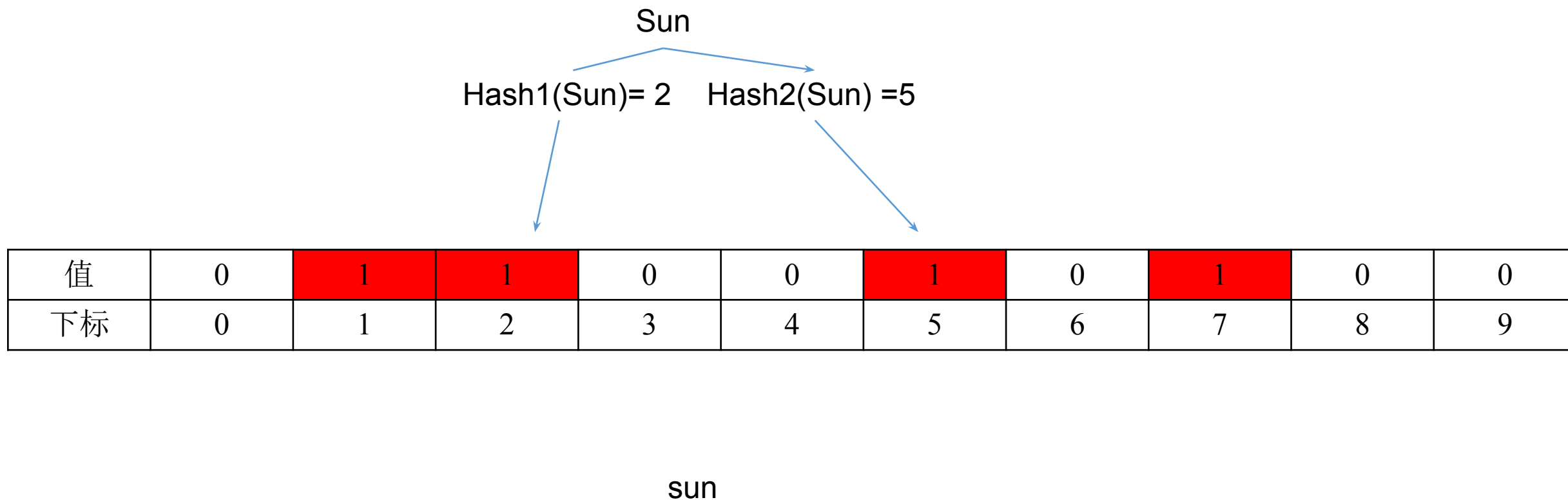
# Bloom Filter 误判率

Bloom Filter Poetry  
False is always False.  
True may be True.

How many false is hidden in true?

## Interview: How to find in bloom filter?

- 如何检查“sun”在bloom filter 里面？



# Bloom Filter精确度跟什么有关？

1. 哈希函数个数
2. 位数组长度
3. 加入的字符串数目

# Bloom Filter 误判率

举个例子如果  
哈希函数的个数15个、  
位数组大小200w、  
加入的字符串数量10w个的话、  
判断2000w个新的字符串  
误判率在 3~4% 左右

计算误判率公式: [https://en.wikipedia.org/wiki/Bloom\\_filter](https://en.wikipedia.org/wiki/Bloom_filter)

# Bloom Filter

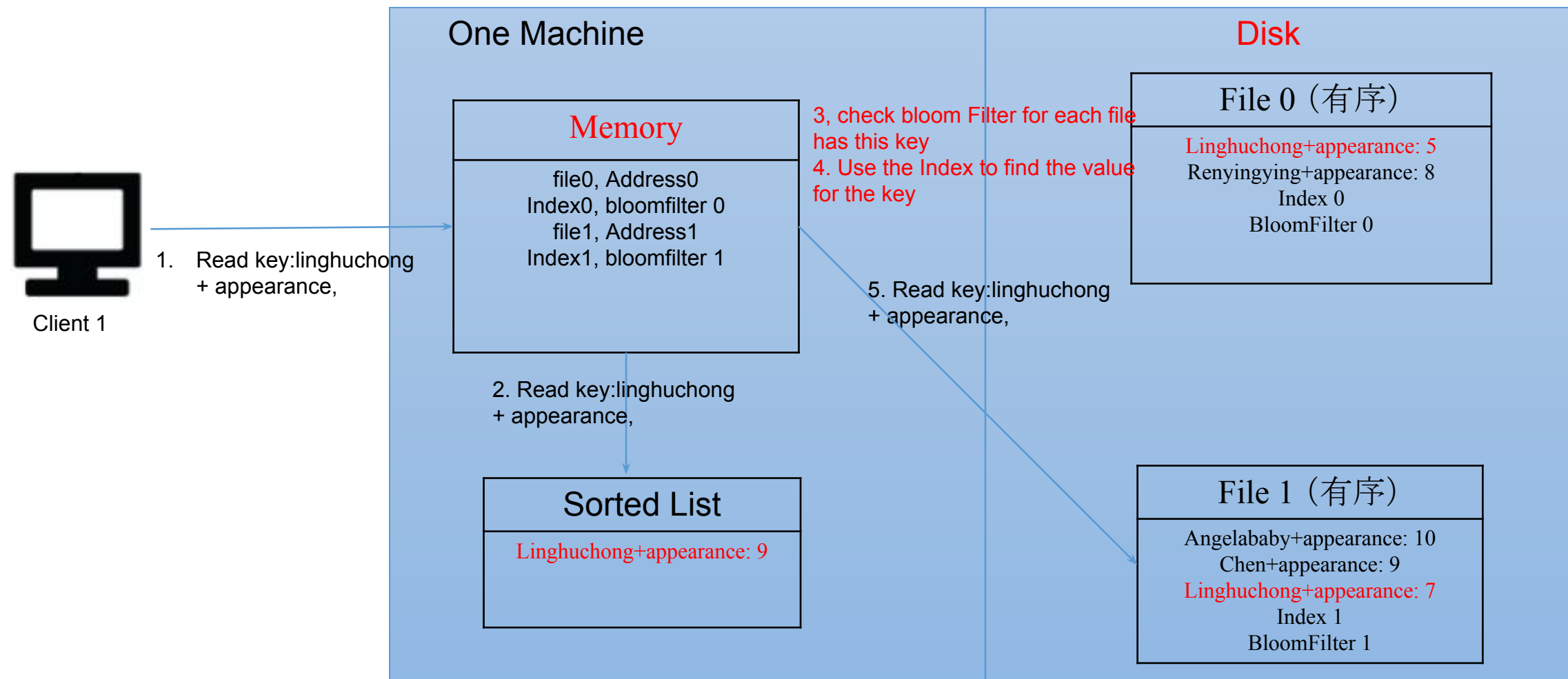
可以高效帮助我们查找key是否在sstable里面

Bloomfilter里面能够找到key的话, 接下来我们再用index去查找

参考阅读:<http://bit.ly/1sUPuwk>

# 完整的读出过程(with Index,BF)





# Specific Name in Bigtable

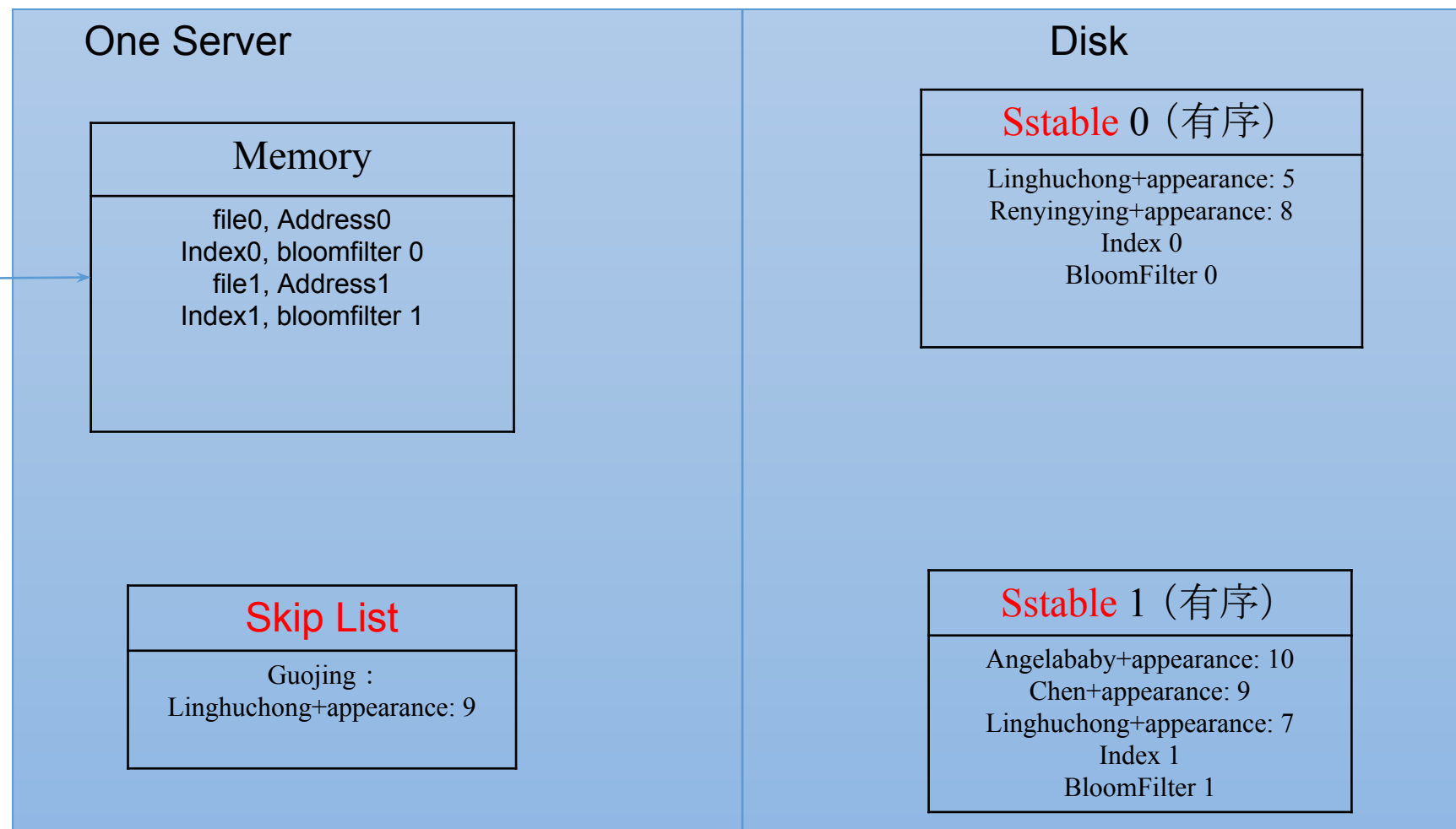
String is Store in the File.

Sstable = Sorted String Table

Sorted List 用 Skip List 实现



Client 1



# 我们已经学会了 一台机器Bigtable的操作

读/写

Key: 令狐冲+颜值

Value: 5

| 姓名  | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |
| 郭靖  | 9  | 180 |
| 东邪  | 7  | 170 |

### 1. Skip List

Code: [https://github.com/petegoodliffe/skip\\_list](https://github.com/petegoodliffe/skip_list)

Wiki: <http://bit.ly/2g0C29a>

### 2. Sstable

Google Sstable Page: <http://bit.ly/1kqwrFe>

Interviewer: How to read/write  
key:value from 1PB file

# Scale

## Sharding?

| 姓名  | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |
| 郭靖  | 9  | 180 |
| 东邪  | 7  | 170 |

# Sharding

Vertical Sharding?

Horizontal Sharding?

难点：取令狐冲“颜值”是不是会取“身高”，“武功”相关属性





|     | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |
| 郭靖  | 9  | 180 |
| 东邪  | 7  | 170 |

Consistent Hash(row key: 姓名)

| 表1  | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |

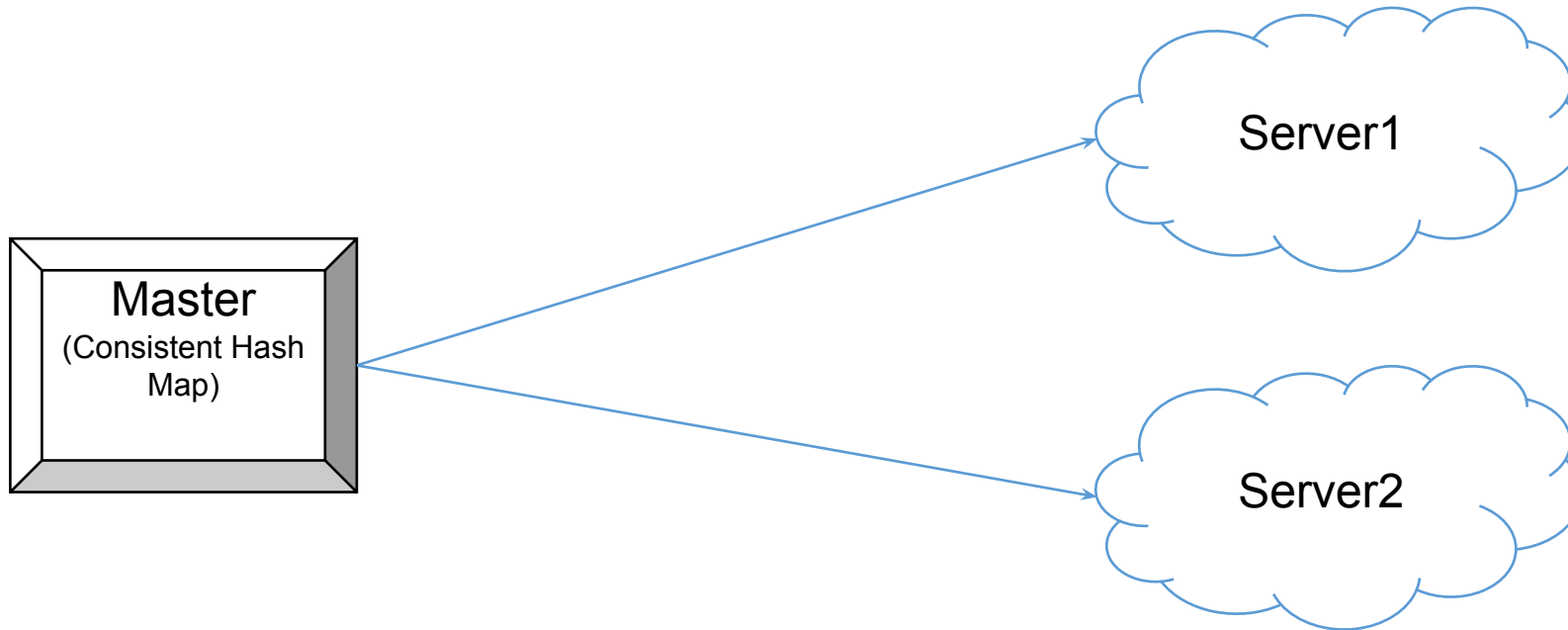
| 表2 | 颜值 | 身高  |
|----|----|-----|
| 郭靖 | 9  | 180 |
| 东邪 | 7  | 170 |

# 一台机器搞不定，那么需要多台机器了

Master + Slave

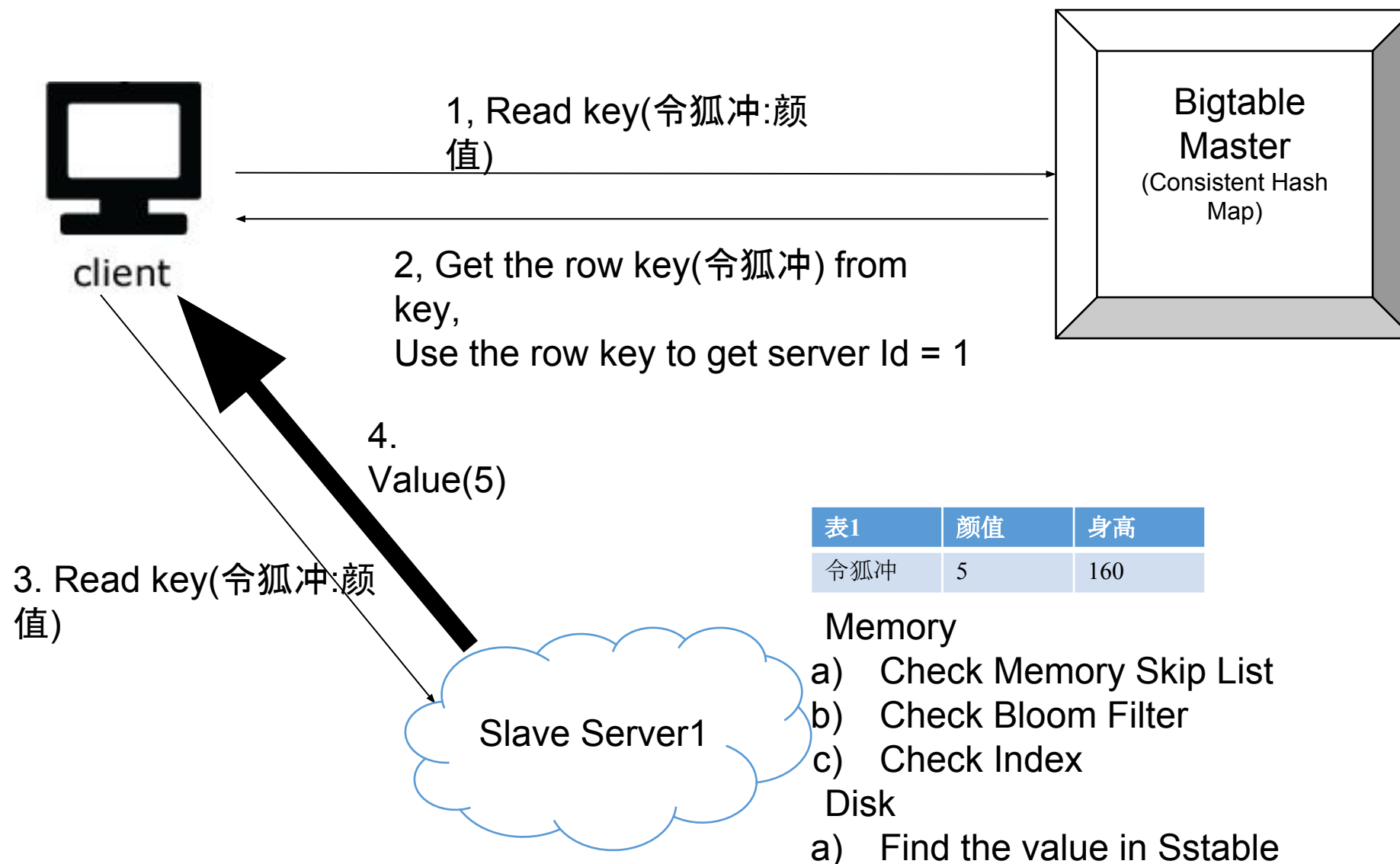
## Interviewer: How to manager server?

- Key
  - Master + Slave
  - Master has HashMap[key, server address]



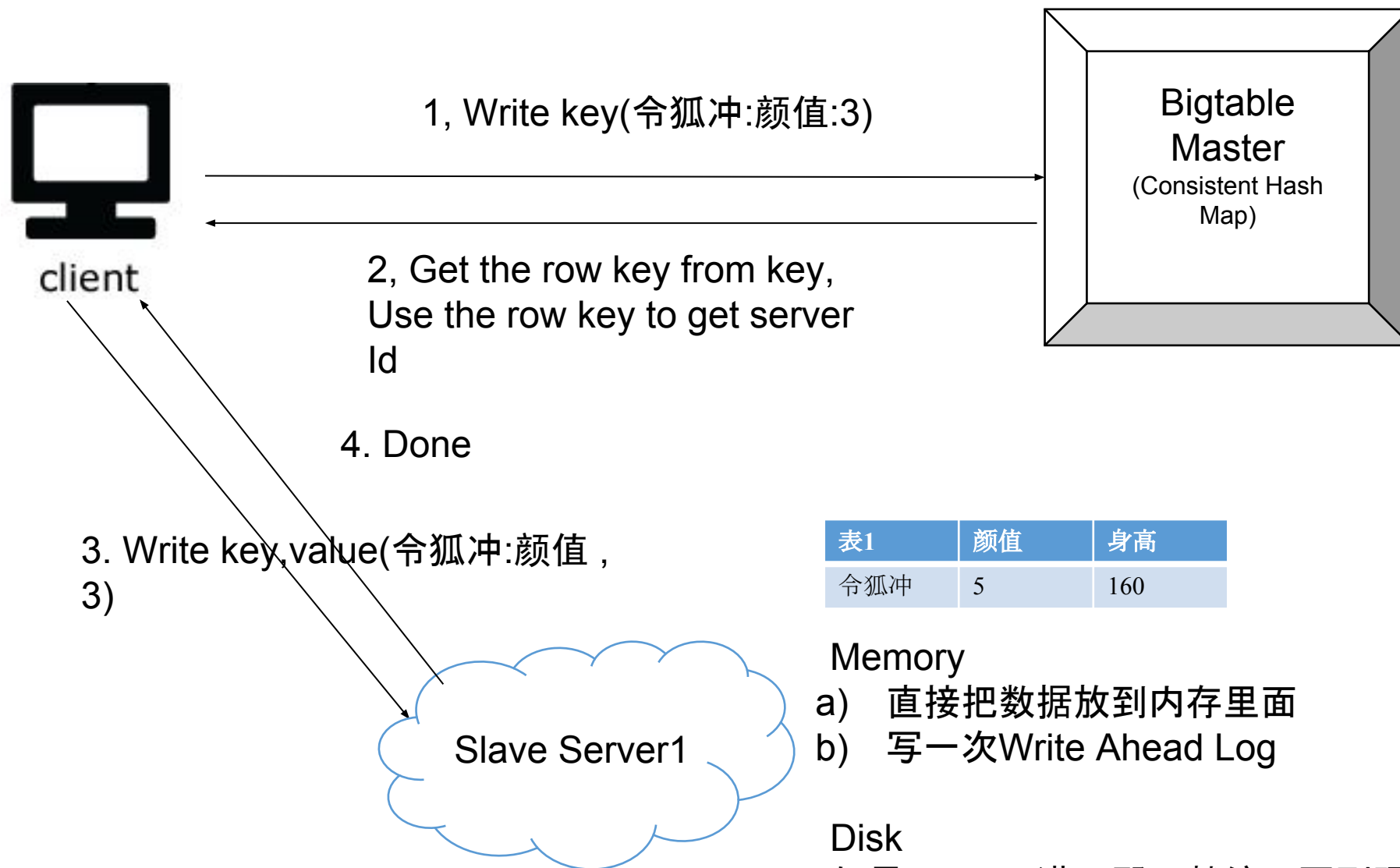
Interviewer: How do we read in  
bigtable with multi-server?

# Interview: How to read a key? (宏观)



Interviewer: How do we write  
bigtable?

## Interview: How to write a key? (宏观)



| 表1  | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |

### Memory

- a) 直接把数据放到内存里面
- b) 写一次Write Ahead Log

### Disk

如果SkipList满了那么就统一写到硬盘



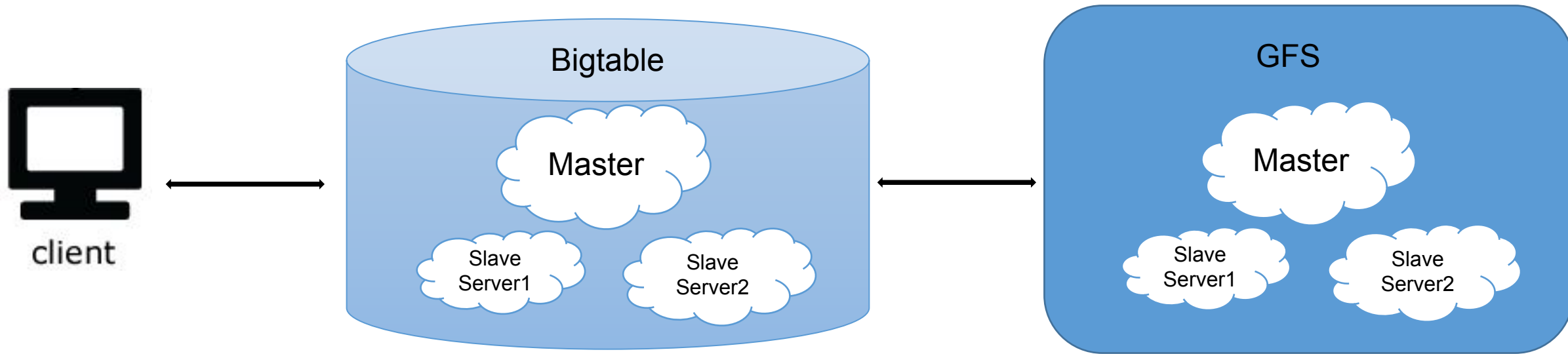
# Interviewer: 每台机器数据越写越多存不下怎么办？

现在所有的数据都存在Slave Server disk里面

# 把所有数据存到GFS里面

Advantage:

1. Disk Size
2. Replica
3. Failure and Recovery



# Bigtable vs GFS

都是Master + Slave

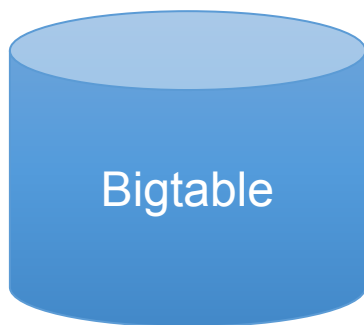
是否就用一个Master+Slave把两个都实现了？

# Sstable 怎么写到GFS里面呢？

难点：Bigtable里面存储单位是Sstable  
GFS读写单位是chunk

# Bigtable Naming

|     | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |
| 郭靖  | 9  | 180 |
| 东邪  | 7  | 170 |



Consistent Hash(row key: 姓名)

| 表1  | 颜值 | 身高  |
|-----|----|-----|
| 令狐冲 | 5  | 160 |

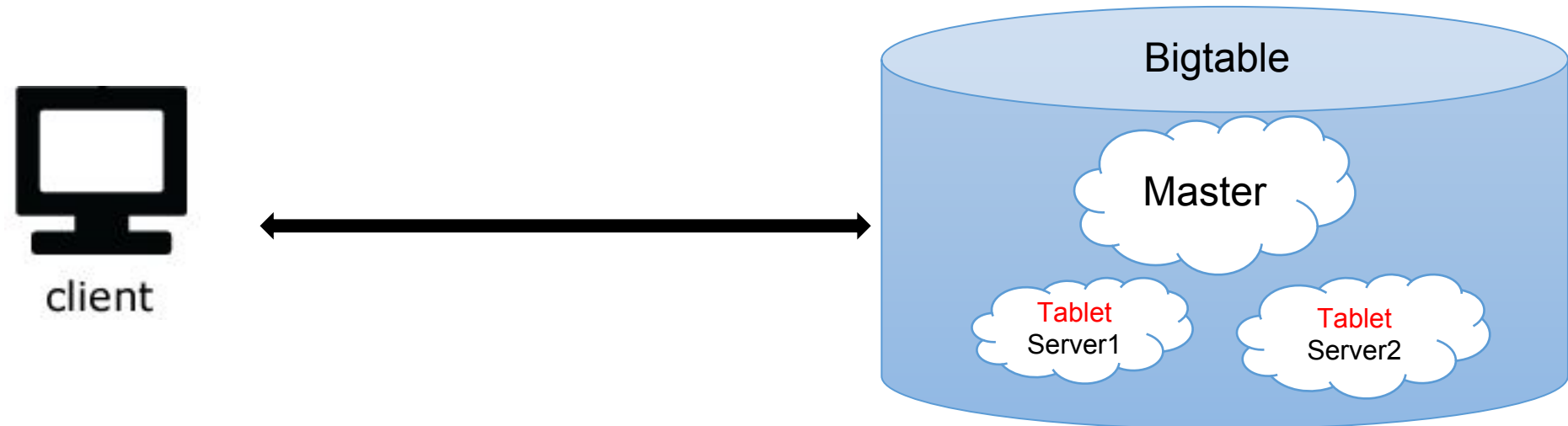
  

| 表2 | 颜值 | 身高  |
|----|----|-----|
| 郭靖 | 9  | 180 |
| 东邪 | 7  | 170 |

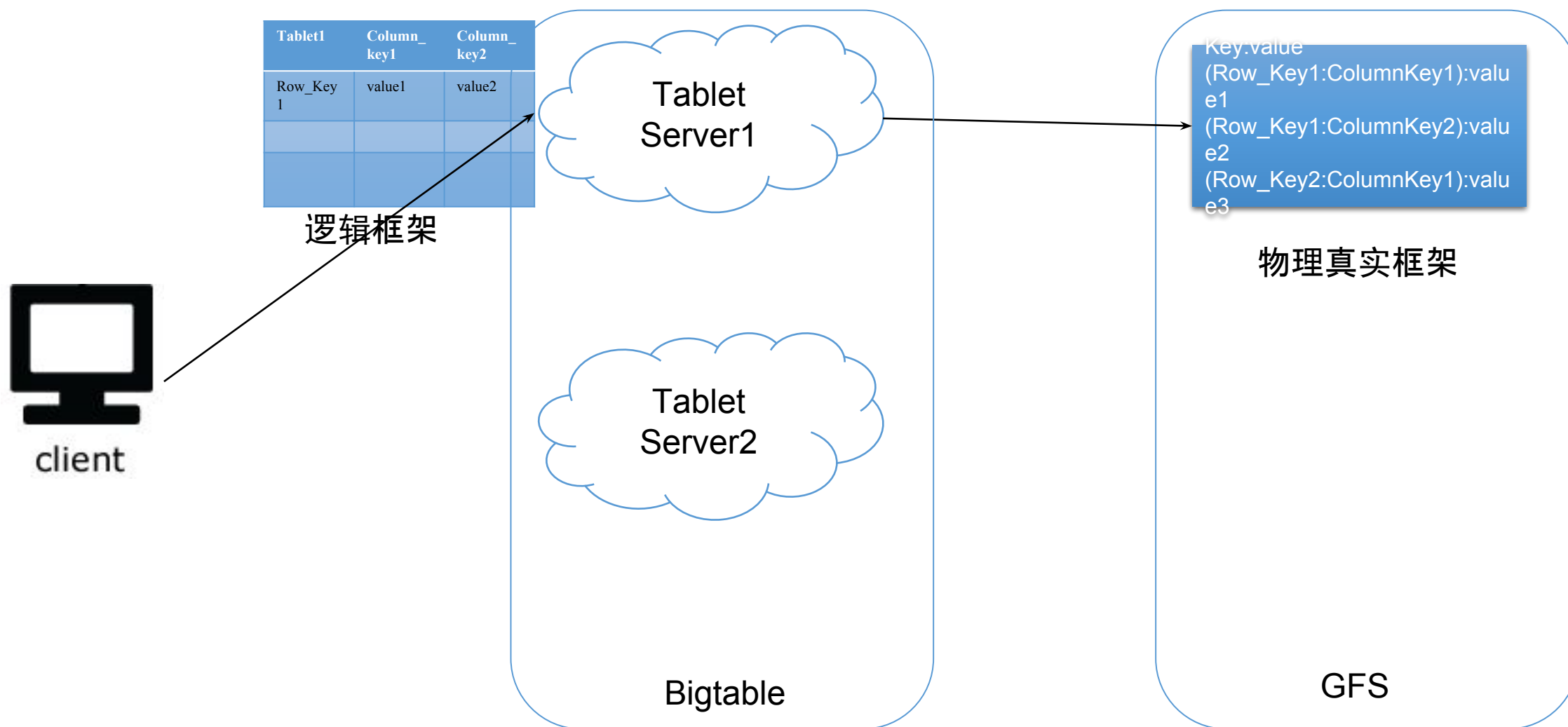


# What is Tablet Server

Tablet Server = Store Tablet Slave Server







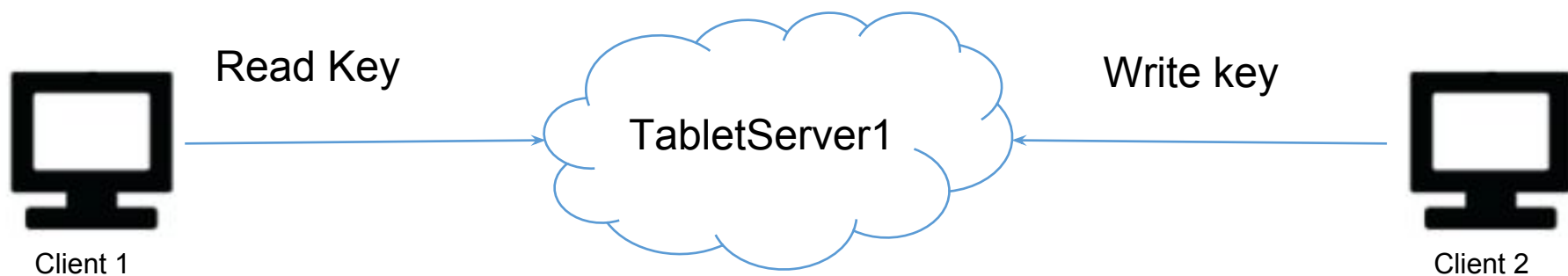
# 看看还有什么问题没有解决

读和写同时发生？

写的过程当中，有读请求？

Race Condition

Interviewer: What if we read and write the same key at same time?



Race Condition

- <http://bit.ly/25FBHM4>



# We need a lock

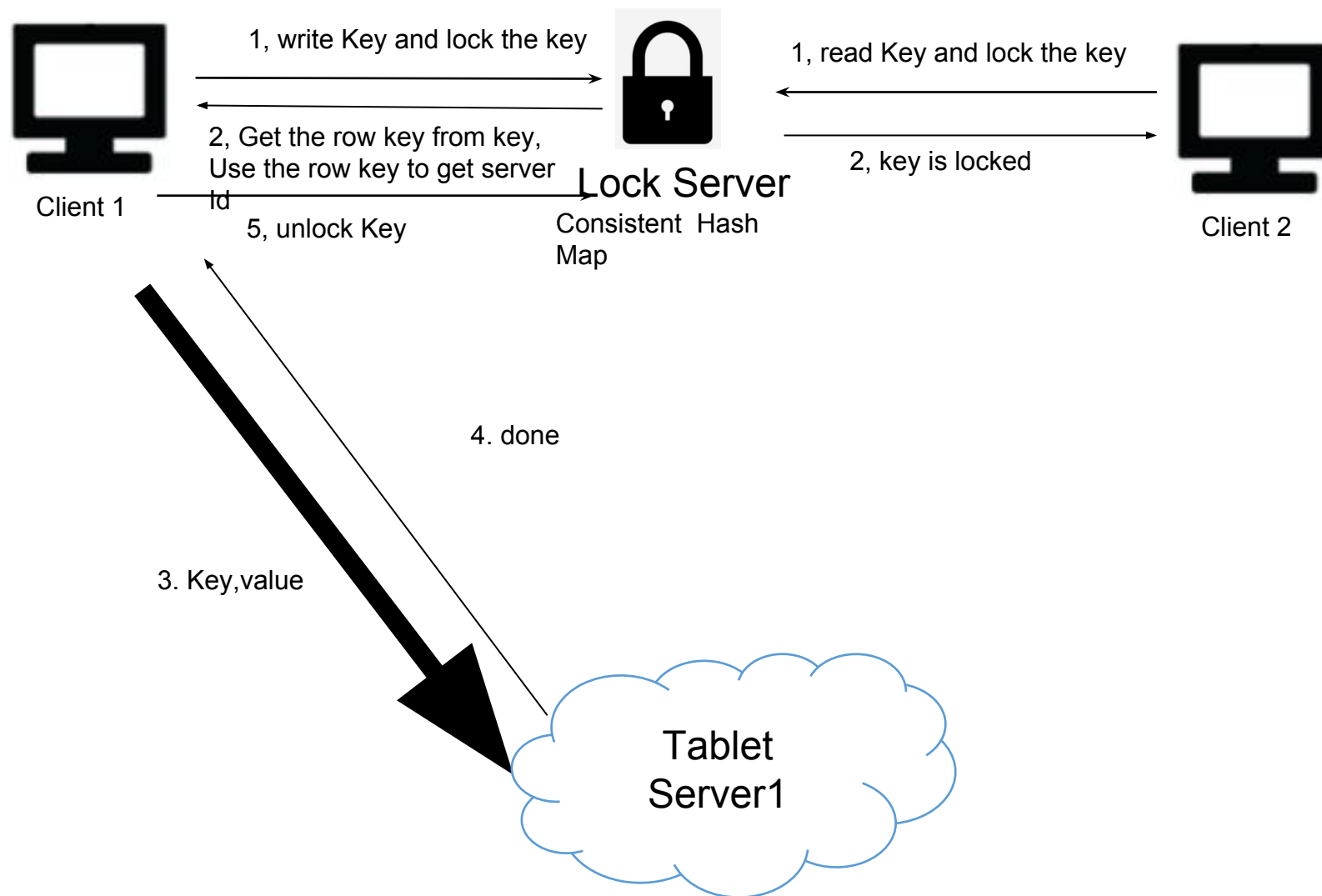
We need a distributed lock.

由多台机器组成的分布式锁服务

- Chubby
- Zookeeper
- Read More:
  - <http://bit.ly/1Pukiyt>
  - <http://bit.ly/1TOWIsR>



# Interview: How to write a key?

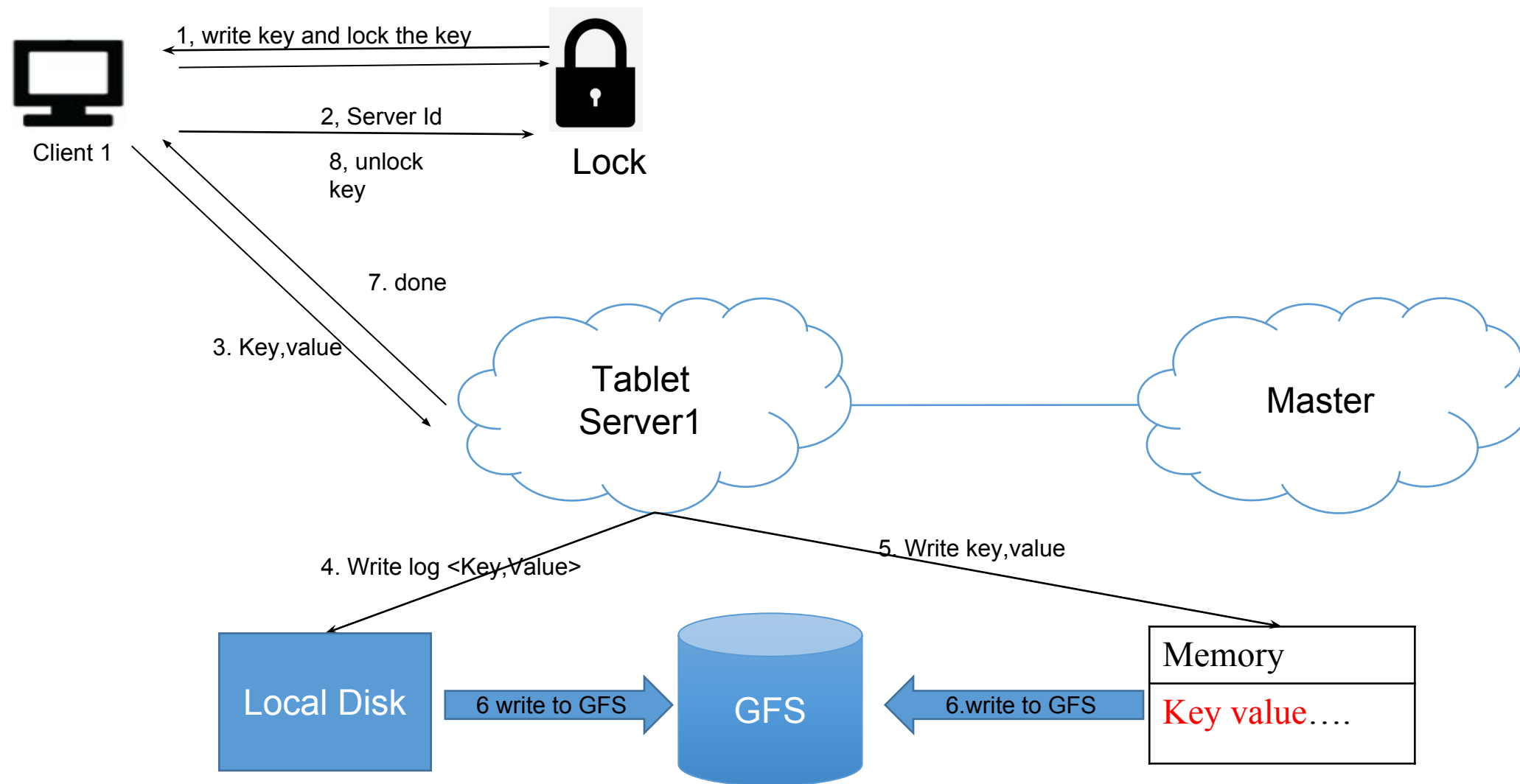


# Advantage Distributed Lock

- The Metadata is store on the Lock
- Lock 本来要存储Metadata那master就不需要存储MetaData了

# Summary of Write

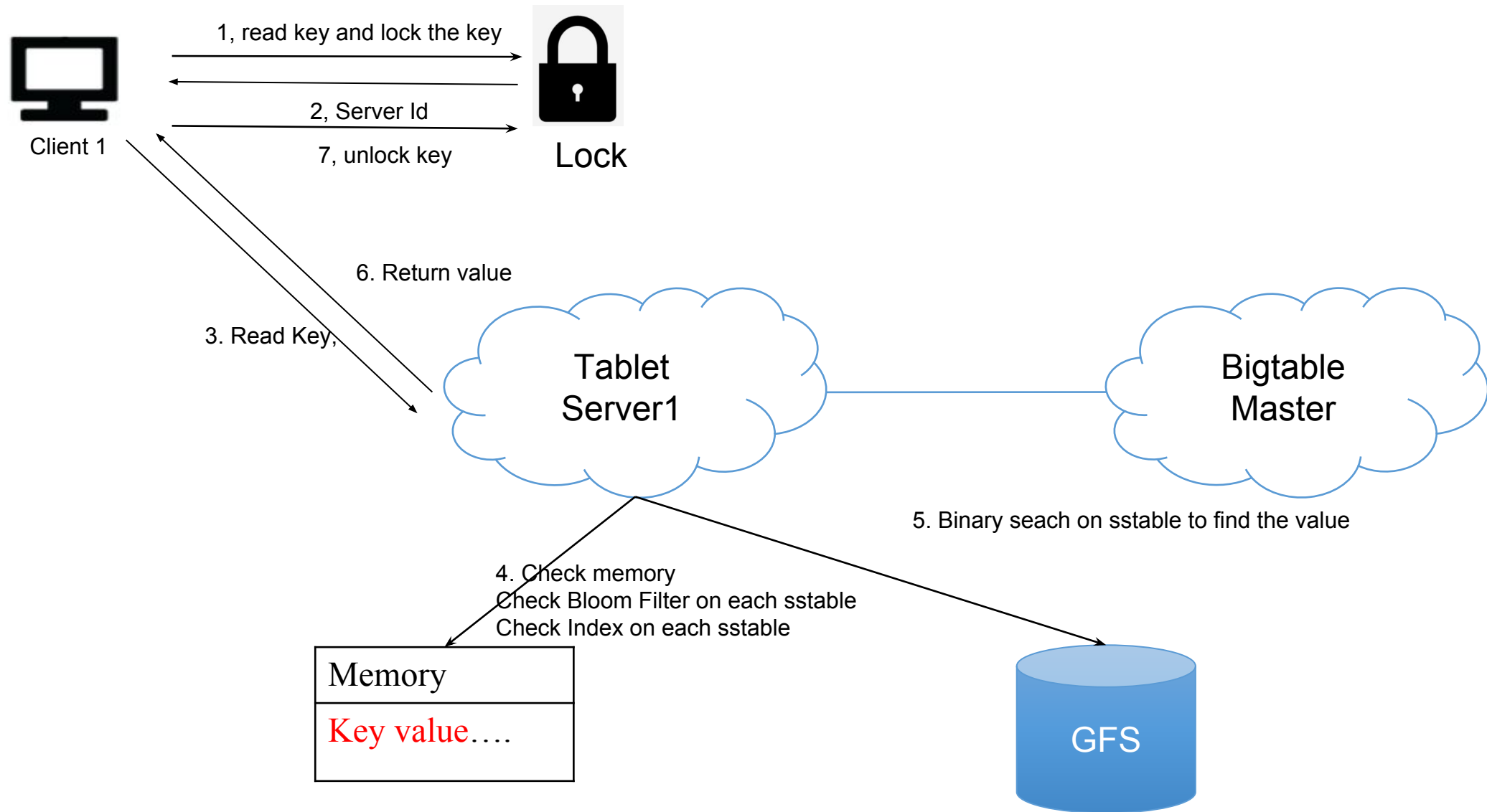
# Summary of write





# Summary of Read

# Summary of Read



- Design
  - Client + Master + Tablet Server + Distributed Lock
- Client
  - Read + Write
- Tablet Server
  - Maintain the Data (Key value pairs)
- Master
  - Shard the file
  - Manage the servers health
- Distributed Lock
  - Update MetaData
  - Maintain the MetaData
  - Lock Key

## Key Point

### Write Optimization

- write append
- Sstable

### Read Optimization

- Binary Search on Disk
- Index
- Bloom filter

### Bigtable

- <http://www.cse.buffalo.edu/~mpetropo/CSE736-SP10/slides/seminar100409b1.pdf>
- <http://www.cs.colostate.edu/~cs435/slides/week11-B.pdf>
- <http://read.seas.harvard.edu/cs261/2011/bigtable.html>
- <http://the-paper-trail.org/blog/bigtable-googles-distributed-data-store/>
- <http://courses.cs.washington.edu/courses/csep552/13sp/lectures/6/bigtable.pdf>

### DevelDb+LSM Tree

- <http://www.xuebuyuan.com/1537388.html>
- <http://www.cnblogs.com/fxjwind/archive/2012/08/14/2638371.html>



