摘要

大数据一词反应了现在社会的趋势。据调查，到2020年每人每秒将产生1.7MB的数据，全球数据量将达到44万亿GB。在如此庞大的数据面前，如何存储变成了我们必须思考的问题，云存储就这样诞生了。

然而当我们放弃本地存储，选择更便捷的云存储的时候，是否考虑过数据的安全问题呢。在实际过程中、云服务并不一定是完全可信的，当用户面临数据泄露、服务器宕机等问题,怎样让用户确认云端数据存储不泄露，所以云存储服务器需要向用户提供完整性证明、还必须支持动态数据操作、同时必须提高通信、计算的性能。

本文针对以上问题：分析了两种数据完整性模型PDP和POR模型，并介绍了基于MHT的数据认证，并实现了基于跳表结构的认证。并给出了跳表结构的时间空间复杂度。本文的主要工作如下：

1. 介绍PDP和POR模型，分析了两种模型优缺点，并给出了支持动态数据操作和公开验证的关键信息。
2. 介绍了实现数据完整性认证的俩个关键的数据结构跳表和二叉树。并介绍了它们的相关操作，同时实现了跳表数据结构，并测试了这种数据结构的性能。
3. 介绍了基于MHT的数据认证，并实现了基于跳表结构的数据认证的两种算法

实现了基于跳表结构的数据认证，并给出了测试结果。

关键词：数据完整性、认证跳表、云存储、数据认证结构

Abstract

The term "big data" reflects the current trend of society. According to the survey, 1.7 MB per person per second will be generated by 2020, and the global data volume will reach 44 trillion GB. In the face of such a huge amount of data, how to store has become a problem we must think about, and cloud storage was born.

However, when we abandon local storage and choose more convenient cloud storage, have we considered the security of data? In the actual process, cloud service is not always completely trusted. When users face problems such as data leakage and server downtime, how to let users confirm that cloud data storage is not leaking? So cloud storage server needs to provide users with integrity proof, also must support dynamic data operation, and must improve the performance of communication and computing.

In this paper, two data integrity models, PDP and POR, are analyzed, data authentication based on MHT is introduced, and authentication based on jump table structure is realized. The time and space complexity of the jump table structure is also given. The main work of this paper is as follows:

1. The PDP and POR models are introduced, their advantages and disadvantages are analyzed, and the key information supporting dynamic data operation and public validation is given.

2. Two key data structure jump tables and binary trees for data integrity authentication are introduced. At the same time, the data structure of jump table is realized, and the performance of the data structure is tested.

3. Introduces data authentication based on MHT, and implements two algorithms of data authentication based on jump table structure.

The data authentication based on jump table structure is realized, and the test results are given.

Key words: data integrity、authentication jump table、 cloud storage、 data authentication structure

课题介绍

## 课题背景

密码学诞生于20世纪70年代，在此之前没有专门形成一门学科，这受益于计算机的发展。

计算机为加密技术提供了新的工具，也让破译者更容易暴力破解。

1949年“保密系统的通信理论”的著名论文问世，将信息引入了密码，激发学者研究密码学的兴趣，奠定了密码学的理论基础。

密码学除了信息的保密之外，还有认证即抵御对手的主动攻击，这在近几年体现的淋漓尽致。

近年来，网络和信息的覆盖越来越广，数据产生已成指数增长趋势。为了存储数据，企业也不得不购买更多的存储设备。这样不仅导致资金的浪费也带来了管理难度。

云存储的概念就诞生了，并逐渐代替了传统的存储方式。各大IT商提出了自己的云服务，比喻阿里云、华为云、百度网盘等等。云服务器宕机、物理机损坏、数据删除等事件的发生，云存储上的数据完整性立刻成为企业关注的要点。

课题意义

云存储的安全面临的问题：

1. 黑客攻击
2. 物理故障
3. 数据损坏不告知用户

云服务器提供的服务不一定是完全可靠的，所以必须提供一个可靠的数据完整性验证的方法。必须满足以下几点：

1. 能公开认证，用户不能实时在线，需要第三方进行周期性检验
2. 可以进行无数次审
3. 防止数据泄露，因第三方参与，必须保证第三方不能知道数据信息。
4. 动态数据操作

国内外研究现状

1989年william Pugh提出了跳表数据结构。

2001年Goodrich在跳表的基础上提出了认证跳表，利用跳表中节点存储数据的hash值，并通过相应的哈希算法实现数据的可信认证

2009年Erway以RSA树和基于认证跳表构造了俩个PDP协议

本文的主要工作

二 **相关基础知识**

随着网络的发展，数据传输的速度也越来越快。密码学的重要性也越来越重要，消息认证是

密码学的一个重要内容。本章详细的对密码学级密码学中消息认证的基础知识做出详细的介绍。具体包括：困难问题、椭圆曲线、认证、Hash函数、以及典型的Hash算法SHA和MD5

## 困难问题

困难问题是密码学的基础，所有的加密算法和认证算法必须依附于困难问题，即没有困难问题就没有密码学。想构造一个密码学算法，必须找到一个与之相对的困难问题。

大整数因数分解问题

给定两个素数p，q，计算比较容易；

给定大整数n，求n的素因数p，q使非常困难。

虽然说大整数因数分解是一个困难问题，但随着计算机的硬件发展，这个困难问题也越来越容易破解。依据这个问题构造的RSA算法现在512bits和1024bits相继与1999年和2009年被破解。现在要保证RSA算法的安全n大约要去2048bits。

离散对数问题

已知有限循环群及其生成元g和阶。

给定整数a，计算元素很容易；

给定元素h，计算整数x，0≤x≤n,使得非常困难。

目前，有限域上的离散对数问题仍没有找到一种方法能有效破解。它的难易程度和生成元的取值无关，只和群中的元素数量有关。一般来说当k取值很小的时候，我们可以利用计算机穷举得到结果，现在的k一般取为1024bits。

椭圆曲线

椭圆曲线是在射影平面上满足方程

（其中a、b、c、d、e为常数）

上所有点的集合。

在诸多困难问题中，椭圆曲线上的离散对数问题要比其他问题难解的多，据研究，160位ECC加密安全性相当于1024位RSA加密，210位ECC加密安全性相当于2048位RSA加密。

椭圆曲线密码体制是比较安全且密钥长度较短的公钥密码体制

## 认证

认=检查、确认。证=身份、证明。认证其实就是为了确认身份、检查权限。

认证包括身份认证和消息认证

身份认证是指验证信息的发送者是否为真，包括信源、信宿等。

消息认证是指验证信息的完整信，验证数据在存储或传输过程中未被篡改、重放等

常见的攻击方式：

1. 重放攻击：偷听者记录当前的消息，在适当的时间再次发送。

对策：增加时间戳

1. 字典攻击：利用常用的口令（例如123456、password等）对消息进行破解。

对策：在口令里面增加额外信息

### Kerberos认证系统

认证系统是防止消息被篡改、重放等攻击的一种有效方式，确保接受者确认消息的真伪。

Kerberos是身份验证协议、是Windows 2000的默认身份验证服务，是基于tcp/ip网络协议设计的可信协议。

步骤如下:

1. client(客户端)发送请求到验证服务器（AS），验证服务器创建一个“session”（根据你的请求），并产生一个代表服务器的随机值。
2. client将这个”session”发送给授权服务器（TGS），TGS会确认”session”，并返回一个消息，并打上时间戳。
3. client向消息服务器发送TGS返回的消息，选择接受或拒绝，并执行相应的服务。

## Hash函数

Hash函数，即散列函数，音译为哈希函数，目的是将任意长的消息压缩为一个固定长度的摘要。

Hash的特点：

1. 输入域中的元素是无穷的
2. 输出域中的元素是有限的
3. 当输入发生微小变化时，输出会发生巨大变化
4. 对于任意输入x计算H(x)是比较容易的，知道H(x)，寻找x是比较困难的

Hash构造方法

1. 直接寻址法

2. 数字分析法

3. 平方取中法

4. 折叠法

5. 随机数法

6. 除留余数法

MD5

MD5 (message-digest algorithm 5)，经由MD2、MD3和MD4发展而来。主要用途：1防止信息被篡改、2防止直接看到明文、3防止抵赖。

MD5原理

**1、数据填充**

对消息进行数据填充，设消息长度为a，a满足(a+x) mod 512=448。可以算出需要填充长度x。填充方法：在消息后面进行填充，填充第一位为1，其余为0。

再把x的长度转换为二进制共64位，如果长度大于那么只取其后64位二进制数。并把这64位填充到消息上，现在消息长度为512的整数倍。

**2、数据初始化**

准备需要用到的数据：

* 4个常数： A = 0x67452301, B = 0x0EFCDAB89, C = 0x98BADCFE, D = 0x10325476;
* 4个函数：F(X,Y,Z)=(X & Y) | ((~X) & Z); G(X,Y,Z)=(X & Z) | (Y & (~Z));  H(X,Y,Z)=X ^ Y ^ Z; I(X,Y,Z)=Y ^ (X | (~Z));

**3、数据分组**

四种操作（Mj表示第j快的32位数据，ti是floor(abs(sin(i)) × 2^32)，<<<s表示循环左移s位）

FF(a,b,c,d,Mj,s,ti)表示a=b+((a+F(b,c,d)+Mj+ti)<<<s)

GG(a,b,c,d,Mj,s,ti)表示a=b+((a+G(b,c,d)+Mj+ti)<<<s)

HH(a,b,c,d,Mj,s,ti)表示a=b+((a+H(b,c,d)+Mj+ti)<<<s)

II(a,b,c,d,Mj,s,ti)表示a=b+((a+I(b,c,d)+Mj+ti)<<<s)

把最终消息以512位为一组，分成若干组，以第一组为例：把这512位分成16快32位

每种操作进行16次，得到最后一组a，b，c，d。

每组操作完成之后A=a+A，B=b+B，C=c+C，D=d+D，进行下组计算，最后将ABCD级联

输出的结果就是MD5码。

SHA-1

SHA-1是由美国标准技术局（NIST）颁布的国家标准，这个算法也是有MD4发展而来。它和MD5的原理基本一样，这里我就只说MD5和SHA的区别。

SHA-1 与 MD5 的比较

l 对强行攻击的安全性：SHA-1得到的结果比MD5码长32 位。使用强行技术， MD5是

量级的操作，而对SHA-1则是量级的操作。这样，SHA-1能保证更大的安全强度。

2 速度：在相同的硬件上，因为SHA-1的循环次数多且缓存区大（160bits：128bits），所以SHA-1 的运行速度比 MD5 慢

可证明安全

一个密码学协议的提出，在没有可证明安全的概念以前，只能经过时间才能得出是否安全。

可证明安全是一种规约方式。如果没有敌手能实现其目标则证明该协议是安全的。

过程为：

描述协议

描述敌手的目标（越低则方案越安全）

描述敌手的能力（越强则方案越安全）

构造安全模型（利用困难问题构造）

归约证明（归约到困难问题不可解上）

例子

1、RSA签名

敌手的目标：

Total Break（完全攻破）

Universal Forgery（对任意签名伪造）

Selective Forgery（可选择对签名伪造）

Existential Forgery（对存在的签名伪造）

Strong Existential Forgery（对多次存在的签名伪造）

敌手的能力：

No Message（没有信息）

Known Message（知道信息）

Generic Chosen Message（知道通用信息）

Directed Chosen Message（知道定向信息）

Adaptive Chosen Message（可以选择性知道信息）

2、RSA加密

敌手的目标：

Total Break（完全攻破）

Decrypting a given Ciphertext（解密给定密文）

Obtaining useful information from Ciphertext（从密文中获取信息）

敌手的能力：

Key Only Attack（唯密钥攻击）

Chosen Plaintext（选择明文攻击）

Generic Chosen Ciphertext（选择密文攻击）

Adaptive Chosen Ciphertext 1（自适应的选择密文攻击）

Adaptive Chosen Ciphertext 2（自适应的选择密文攻击2）

在Adaptive Chosen Message情况下，可以抵抗Existential Forgery。说明此协议是安全的

三 云存储的完整性验证

云存储作为一种线上的存储方式，我们肯定会怀疑存储的数据是否安全，是否被别人修改，数据认证成为了云存储中的一个重点研究对象。本章简述了云存储中数据认证的重要的数据结构Merkle Tree，以及典型的数据认证算法。

数据完整性方案模型：

在现有的数据完整性认证模型中：PDP和POR是最典型两种模型。这两种模型的出现为以后的数据认证模型奠定了基础。

PDP模型高效，因为它不需要对数据进行加密。

POR可以恢复以损坏的文件。

PDP模型

Ateniese等人在2007年提出了PDP模型，它是一个基于概率统计的一个模型，是采用抽样的形式，通过一个或几个子文件来判断整个文件的完整性。客户端在做预处理时，利用少量的元数据去向服务器发起挑战。

图片

PDP包括以下算法

KeyGen() ->(pk,sk):客户端生成公钥pk=（N,g）和私钥sk=（e，d，v）其中e是事先说明的大数e>且d>，v，输出pk，sk。

TagBlock(pk,sk,m,i) ->:客户端利用pk和sk生成，。输出，称为元数据的标签。

GenProof(pk,F,chal,∑) –>V:服务端利用输入的pk和chal，遍历所有的元数据，得到V=(T,)

其中T=，H(g mod N),输出V；

CheckProof(pk,sk,chal,V)->{success,failure}:客户端利用pk和sk和chal生成，比较V和是否相等，若相等则输出success，否则failure。

我们生成PDP协议和规范分为两步：setup和challenge。

setup：客户端运行KeyGen()生成公钥pk=（N,g）和私钥sk=（e，d，v）并在客户端上存储。将原始文件分块，再利用TagBlock(pk,sk,m,i)生成，并将原始文件和集合{}一同发送个服务器。

Challenge：客户端以n为单位按周期向服务器发起挑战，即在原始文件中随机抽取c个样本组成chal集合，并chal请求发送个服务器。服务器利用GenProof(pk,F,chal,∑)生成证明V，并返回客户端。收到V，CheckProof(pk,sk,chal,V) 来确定原始文件是否完整。

POR模型

Juels等人于同年提出了POR模型的概念。POR和PDP一样都可以保证数据的完整性，但是POR模型设置了客户端发送挑战的次数的最大值，导致了POR模型只能支持静态数据验证和恢复。

图片

POR包括以下算法

KeyGen()->K

Encode(F;K,) ()->()

Extract(;K,) ()->F

Challenge(;K,) ()->c

Respond(c,)->r

Verify((r,); K,)->b

POR算法和PDP算法执行基本一致，只有

Encode(F;K,) ()->()算法中对文件进行了加密，并在文件中加入”哨兵”数据和原始数据混合生成新的数据。

Extract(;K,) ()->F算法是通过正确的验证来返回原始数据，由于”哨兵”的产生是随机的有限的，每次数据恢复就会返回一个”哨兵”，从而导致只能做有限次验证。

具体的POR算法请参考文献。

公开验证

在本地存储的文件，我们一般通过访问全部数据来验证完整性，如果在云存储中也采用同样的方式进行数据完整性的校验，会耗费计算机大量的资源。即使不考虑计算机的性能问题，我们也不能每时每刻的让客户端保持连接状态。因此，我们把数据完整性校验委托给第三方来代替我们进行验证工作。

在公开验证中，引入了一个第三审计员，在客户端与服务端建立可信机制，所以又把这种验证方式称为数据审计。

公开验证的关键：

在PDP和POR协议中，只有客户端拥有原始数据的私钥sk，其他人无法进行验证。同时这两种算法的客户端计算复杂度、存储开销、通信开销都比较高。如果提高验证效率和扩展算法得以支持公开验证成了学者关注的主题。现在的方案大多都是基于同态验证标签（HVTs）来实现的。

同态验证标签具有两个基本性质：

1. 无数据块验证：只需要提供原始数据由标签构造的证据，就能进行数据审计
2. 同态性：已知数据块m，n它们分别对应的同态标签，。数据块f=m||n。

可以通过，来表示数据块f的同态标签。

动态数据更新

实际上,当公司用到云存储来保存文件，这些文件的数量是很庞大的，并且需要随时随地的进行数更新，因此云存储应该支持动态的数据更新。当信息变化时，我们也应该保证这些数据的完整。由于网络资源有限、下载整个文件然后更新，显然是不现实的，所以必须找到一个合理的算法保证云存储服务器进行了数据更新。

在PDP协议的基础上增加三个算法

PrepareUpdate(pk,sk,data)->info:由客户端执行，输入公钥pk，私钥sk，以及要更新的信息data，输出请求信息info。（info包括增删改或查操作和数据信息）

PerformUpdate(info,F,)->{,,}:由服务器执行，输入info，文件F，元数据 ，输出新的文件,新的元数据,更新标识。

VerifyUpadte()->{accept,reject}:由客户端执行（如果是支持公开验证，可以委托给第三方代理执行），输入公钥pk和更新标志，如果验证通过则输出accept，否则更新失败输出reject。

四 跳表和二叉树

跳表和二叉树是两种常用的数据结构，在动态数据更新和数据认证中起着很大的作用。本章主要介绍跳表和二叉树这两种数据结构的原理，以及一些重要的节点操作。

跳表

跳表是一种比较重要的数据结构，在数据认证中起着很大的作用。

### 为什么选择跳表？

在云存储中我们经常要存储大量的数据文件。如果我们直接用二叉树或者链表用来记录文件节点，会影响我们的查询效率。

然而 在目前经常使用的平衡二叉树结构中，它是一种严格的平衡，虽然它的最差的查询时间一定是O(logn)，但它在插入方面需要旋转和overhead等操作。

跳跃表（Skip List）于1987年诞生，其查询 、插入和删除操作的预期时间复杂度为O（logn）。虽然跳表的最差查询时间是不稳定的，但跳表对插入更快，对并发的处理更友好。

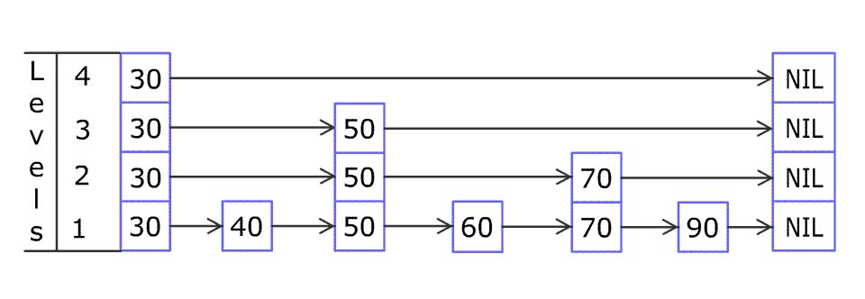
## 跳表的基本原理

跳跃表由多条链构成（，， ……，），满足一下条件：

最底层链表包含所有的元素，链表中的元素升序

每个链表有两个特殊节点 -∞(起始节点)和+∞(终止节点)

S0包含所有元素，对任意的i>0， +1是的子集，且中的每一个节点，有一个下行指针，指向中与之有相同索引值的节点。

 （图片换）

跳表的特点

1. 节点重排，以及顺序性的访问方式
2. 灵活的增删节点

跳表的相关操作：

查询节点：（如果当前节点存在，则返回当前节点。如果不存在则返回离此节点最近的）

1. 从最上层的链（Sh）的开头开始
2. 假设当前位置为p，它向右指向的节点为q（p与q不一定相邻），且q的值为y。将y与x作比较

(1) x>y 从p向右移动到q的位置

(2) x=<y 从p向下移动一格

3. 当p在底链（S0）中且x=<y 输出当前p节点

删除节点

1. 通过查询 判断节点是否存在

2. 存在 则将该节点所在整列从表中删除

3. 将多余的“空链”删除

增加节点

高度的生成方法

（1）level = 0，

（2）产生一个0到1的随机数r，如果r小于一个常数p（），

（3）则执行方案level=level+1且重新执行（2）（3）

（4）否则，输出level。

1. 通过查询 判断节点是否存在
2. 在当前节点p后增加高度为level的节点

跳表的性能

空间复杂度： O(n) （期望）

相关操作的时间复杂度：

查找： O(logn) （期望）

插入： O(logn) （期望）

删除： O(logn) （期望）

空间复杂度分析：

令由高度的生成方法，每个元素插入到第i层（Si）的概率为，则在第i层插入的期望元素个数为n，跳跃表的元素期望个数为 ，当的数时，次数总和小于2n。

S = 

当

S = 2n

空间复杂度为O(n)

实际测试： 共执行了次 maxLevel=3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| P | 平均插入时间 | 平均高度 | 总结点数 | 平均查找时间 |
|  | 0.53976 | 1.7503845 | 17503845 | 0.28027 |
|  | 0.46651 | 1.5754064 | 15754064 | 0.25341 |
|  | 0.60256 | 1.3146518 | 13146518 | 0.31451 |
|  | 0.78464 | 1.1734221 | 11734221 | 0.35154 |

一般我们把p取，此时的平均查询时间最短，如果有要求空间的话，我们也可以退而求其次取p=。这样节点就不会占据大量空间。

二叉树

树的相关定义

结点：树的结点包含一个数据元素以及若干指向其子树的分支

深度：树中结点的最大层次

高度：从叶子结点开始定义，叶子结点为第一层，往上依次递增，直至根结点。

度：结点所拥有的子树数量(二叉树的度最大为2)

终端节点：度为0的结点称为叶子结点

**先序遍历**

即遍历顺序为“根—>左->右”。

**中序遍历**

即遍历顺序为“左—>根—>右”，

**后序遍历**

即遍历顺序为“左—>右—>根”。

二叉树的相关操作：

查询（树不为空）：

1. 要查找的value与当前节点curvalue比较

2. value<curvalue:当前节点移动到左子树，并重新执行1

value>curvalue:当前节点移动到右子树，并重新执行1

value=curvalue：return 当前节点

删除

执行查询，如果不存在，则返回, 否则要删除的结点可能分下面四种情况：  
a. 要删除的结点无孩子结点：直接删除该结点。  
b. 要删除的结点只有左孩子结点：删除该结点且使被删除节点的双亲结点指向 被删除节点的左孩子结点  
c. 要删除的结点只有右孩子结点：删除该结点且使被删除节点的双亲结点指向被删除结点的右孩子结点  
d. 要删除的结点有左、右孩子结点：在它的右子树中寻找中序下的第一个结点，用它的值填补到被删除节点中

插入

1. 树若为空，直接插入
2. 树不为空，执行查找，找到终端节点
3. 插入新节点

五 认证跳表

上章介绍跳表和Merkle Tree这两种数据结构。它们比一般的数据验证算法更适用于云存储的数据验证，它们可以有效的实现对数据的增删查改。本章介绍了基于跳表和Merkle Tree结构的数据认证并给出基于跳表认证的实现方式。

基于跳表的认证原理

对算法中出现的名词的定义：

路径栈：在遍历中经过的节点，并且用栈进行保存。本文用skipNodeStack

来表示。

节点层数：跳表的当前层数。用level表示

isNextNode：查询的节点值是否是大于右节点的值

isPreNode：查询的节点值是否是大于左节点的值

Roberto提出的认证跳表算法有两步，首先获取节点的Hash值，然后根据认证证据和节点关系并与节点的Hash值进行比较。

遍历的路径栈具体步骤如下：

1. 如果当前节点的右节点不为空，则移动到右节点。
2. 如果当前节点处在level=0的地方则将该节点放入isNextNode集合中，否则将该节点的下层节点添加到isNextNode集合中。然后向上遍历，直至到当前节点的最高层。
3. 判断节点的最高层的右节点是否为空，若空则停止计算，否则重1开始。

节点的验证：其实就是用isNextNode和isPreNode来检验遍历的路径栈是否正确。

节点的验证过程步骤如下：

1. 如果isNextNode不为空：

如果当前节点不是最高层，则下个节点处在索引为i的level+1层

如果是最高层，则下个节点处在索引为i+的level层

1. 如果isNextNode为空：

如果isPreNode不为空，则下个节点处在索引为i的level+1层

如果isPreNode为空，则下个节点就是节点

Hash值的计算：

1. 若isPreNode不为空：

Hash(node)=Hash(currentNode)||Hash(preNode)；

2. 若isPreNode为空

Hash(node)= Hash(currentNode)

最后将Hash值和路径栈一并返回，客户端进行验证，若一至则成功，否则失败。

基于跳表的认证实现

开发环境Eclipse+jdk1.8

基于跳表的认证是通过修改跳表的节点的指针，形成一个伪二叉树结构。同时又保留了跳表的一些特点，比如随机性、有序性。

首先建立跳表，跳表节点的结构如下：

{

Long key;//节点的key

Object value;//节点原始的Hash值

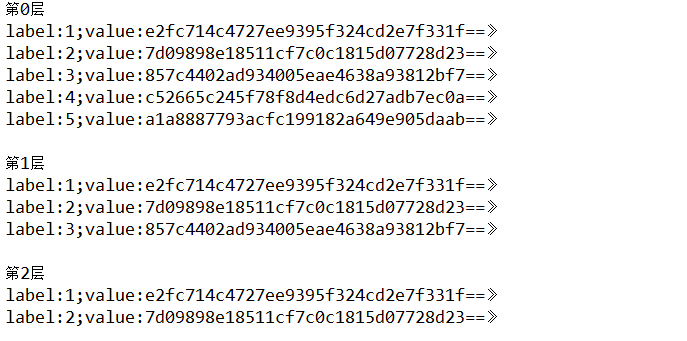
SkipListNode[] forword;//用于存储前节点

SkipListNode[] back;//用于存储后节点

}

跳表的增删改查上述已介绍，这就不具体说明。

然后将跳表加工成认证跳表,我将X:\\hash下的文件利用MD5消息摘要算出Hash值。跳表结构如下



每个节点存储的Hash值分别对应的是5个文件。任意文件的修改都会改变level=0的节点的Hash值从而影响到最后生成的Hsah。

最后的Hash值是 fffffdffffffffff95f324cd2e7f331f

修改文件，然后得到的Hash值为：fffefdffffffffbf95f324cd2e7f331f

基于Merkle Tree的认证

基于Merkle Tree的认证是由Ralph提出的，这个方法的主要思想：

Merkle Tree

Merkle Tree中文名字默克尔树。

