doi:10.3969/j.issn.1000-436x.2012.z2.028

云存储中的数据持有性证明研究综述

付伟，叶清，陈泽茂，吴晓平
（海军工程大学 信息安全系，湖北 武汉 430033）

摘 要：数据的持有性证明允许用户随时知道其数据是否仍然有效地保存在云存储平台中，以及是否可以随时、随地获取到该数据，这是云存储安全中的一个重要的挑战性问题。介绍了数据持有性证明的模型和衡量指标体系，分析了 3 种证明方案，并全面比较了 6 种常见的方法，最后给出了未来需要注意的研究方向。
关键词：云计算；云存储；云安全；数据持有性证明
中图分类号：TP302.1 **文献标识码**：A **文章编号**：1000-436X(2012)Z2-0201-06

Survey of data possession provability proving on cloud storage

FU Wei, YE Qing, CHEN Ze-mao, WU Xiao-ping
(Department of Information Security, Navy University of Engineering, Wuhan 430033, China)

Abstract: Data possession provability allows cloud users to verify whether their data are still integrally preserved in Cloud Storage system, and whether they can obtain the data at anytime, anywhere. It is a challenging research problem of cloud storage security. A data possession proving model and its estimating criteria system were established. Three famous proving schemes were analyzed in detail. Six proving methods were compared from different aspects. Finally, some valuable future research issues and suggestions were pointed out as conclusion.
Key words: cloud computing; cloud storage; cloud security; data possession provability

1 引言

云计算是当前信息技术领域的热点问题之一，代表了 IT 领域向集约化、规模化与专业化发展的趋势[1]。它是继网格计算[2]之后分布式计算技术的又一次重大发展。云计算描述了对组成计算、网络、信息和存储等资源池的各种服务、应用、信息和基础设施等各种组件的一种全新使用模式[3]。

随着信息技术的不断发展，数据容量呈爆炸式增长，人们对于数据存储的要求也越来越高。研究表明，管理数据的成本是获取该数据成本的 2~3 倍。特别是在云计算技术兴起的背景下，针对数据密集型应用出现一类特殊的云计算平台。它们通过集群应用、网格技术或分布式文件系统等功能，将网络中大量各种不同类型的存储设备通过应用软件集合起来协同工作，共同完成复杂的数据管理功能，专门对外提供在线的数据存储和业务访问功能，一般称其为云存储[4]。云存储可以方便地根据需求调整系统规模，能够快速配置/重新配置、部署、提供存储服务和解散存储资源，为用户提供一种按需分配的存储消费模式；云存储服务提供商则依据存储容量、存储时间、访问带宽等指标向用户收取费用。这其中的典型代表包括 Amazon 推出的简单存储服务（S3, simple storage service）[5]平台和 EMC 公司的 Atmos 云存储平台等。

云存储技术极大地节约了社会的资源与能源。

收稿日期：2012-07-10
基金项目：国家自然科学基金资助项目（61100042, 71171198）；湖北省自然科学基金资助项目（2011CDB052, 2012FFB06901）
Foundation Items: The National Natural Science Foundation of China (61100042, 71171198); The Natural Science Foundation of Hubei Province (2011CDB052, 2012FFB06901)

但是云存储还存在众多问题亟待解决，这些问题导致云存储仍不能得到用户的广泛认可与使用。其中一个挑战性问题就是数据的持有性证明问题[6]。在云存储平台中，用户将数据以外包（outsourcing）的方式存放在云存储系统中。在这种新型模式下，用户与云存储服务提供商之间的关系从传统的“服务器/客户端”演变成为“商家/顾客”的关系。从服务方来看，如果没有相应的检测和监督机制，出于规避商业或者法律法规风险的考虑，云存储服务提供商在自身出现问题导致数据丢失的时候会选择隐瞒不报，或者简单地推卸应付的责任。而从使用方来看，用户必须随时知道自己的数据是否仍然有效地保存在云存储平台中，以及自己是否可以随时、随地获取到该数据。只有当存储服务商对用户提供数据是否有效存在、能否成功获取的验证方法，并不断提高服务水平，云存储才能够获得广泛的应用。为此，科研人员提出数据持有性证明的问题，研究如何通过高置信度的技术手段保证存储服务商忠实地按照服务契约维护数据的可用性及完整性，监督其无法规避应付的责任。

本文系统、全面地总结了现有的各种数据持有性证明方法，并对现有方案进行综合比较，分析各自的优缺点。文章的组织结构如下：第 2 节给出模型与衡量体系标准，第 3 节分别讨论基于对称密码学、非对称密码学和基于第三方审计的证明方案，并在第 4 节中对上述方案进行分析比较，第 5 节总结全文，并指出未来需要注意的研究方向。

2 数据持有性证明模型与衡量指标体系

传统存储系统的数据持有性证明主要采用基于访问的方法，例如一些在线存储系统[7]、海量存储系统[8]以及数据库存储系统[9]等。这些系统将数据下载到本地，由用户逐一验证数据的存在性及完整性。这种方式需要频繁地访问服务器上的数据，增加了服务器的负担，严重浪费了网络的带宽资源。在具有数据量较大的应用特征的云存储环境下，这种方式显然效率太低，不适合云存储应用，下文也不再继续讨论。

针对传统方法存在的弊端，研究人员提出另外一种基于“挑战–应答”方式的持有性验证机制[10~17]。在这种机制下，客户端有选择性地提出挑战其中的某一些数据块，服务器根据客户端的挑战要求生成相应数据完整性的证据并发送给客户端，最后由客户端来判断结果。本文提出适用于基于“挑战–应答”方式的一般证明模型如图 1 所示。

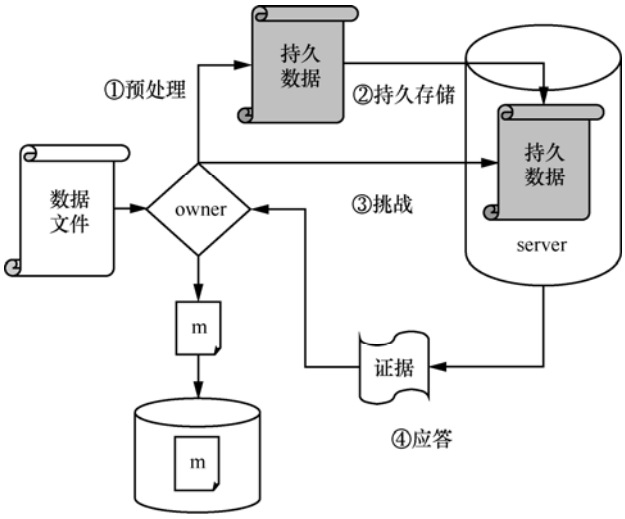

图 1 数据持有性证明模型

由图 1 可见，在该模型中只涉及到两种角色，即数据拥有者（owner）和存储服务器（server）。主要包括 2 个阶段 4 个步骤。

1）数据存储阶段(包括图 1 中第①、②步)

数据拥有者将其数据文件进行一定的预处理过程，例如将数据分片、对分片后的数据进行加密、从分片数据中抽取一些抽样信息等，得到处理后的持久数据文件和用于之后验证的元数据 m；然后将持久数据文件上传给服务端长期保存，而验证用抽样元数据由客户端自己秘密保存。

2）数据验证阶段(包括图 1 中第③、④步)

在数据的生存期内，数据拥有者向服务端提出数据持有性验证挑战，服务端依据拥有者的要求，抽取一些特定的持久数据进行运算处理，获得可证明该数据存在的一个证据，然后将该证据传回给拥有者；拥有者将其与之前保留的抽样元数据 m 进行一些特定的处理，最后根据处理结果判断原始数据是否正确地存储于服务端。

与传统存储系统的存在性证明模型相比，该模型在挑战阶段只需发送少量（一般只有几个字节）的挑战信息；而在服务器应答阶段，可以充分利用强大的云计算能力生成证据。一般的证据信息只有几十、最多几百个字节，因此可以大幅地节约网络带宽资源，特别适合于聚合计算能力突出而数据量庞大的云存储应用。

现有的各种不同方案的区别在于它们对文件

的预处理过程不同，以及计算和验证证据的算法不同。一般来说，衡量一种方案优劣的指标主要包括以下几个方面。

1) 计算开销：包括拥有者对数据文件进行预处理所需的计算开销、服务端计算获得证据所需的计算开销，以及客户端验证证据所需的计算开销；

2) 传输开销：数据拥有者与服务器之间的数据传输量；

3) 存储开销：拥有者存储相关元数据 m 所需的存储开销，以及处理后所得持久数据与原始文件的比例；

4) 更新操作：即是否支持数据所有者对数据进行更新，以及支持何种类型的更新操作，包括 Append、Modify、Delete 以及 Insert 等；

5) 挑战次数：允许用户提出挑战的次数；

6) 可恢复性：部分数据出错时能否恢复原始数据的特性。

3　现有数据持有性证明方法

3.1　基于对称加密体制的方案

基于对称密码学的方案以 RSA 公司的 Juels 和 EMC 公司的 Kaliski 提出的基于岗哨的可恢复证明系统 POR[10]为典型代表。其基本思想是首先用对称加密体制将文件加密并用纠错码编码，然后在编码后的文件中的一些随机位置插入和文件数据不可区分的“岗哨”（sentinel）；检查者在挑战时要求服务器返回一些随机位置的岗哨。作者证明只要服务器以大于一定值的概率做出有效的应答，则文件是可以恢复的。这种方法具有计算开销小的优点，但其缺点在于每次需要消耗掉一个岗哨，因此只能执行有限次的挑战。此外，如果文件需要更新时，这种方案需要找出所有未经使用的岗哨，然后重新编码、重新插入文件中，效率较低。

同属 RSA 实验室的 Bowers 等人[11]在 Juels 等研究的基础上提出了一个 POR 的理论框架，用于改进已有 POR 方案，实现更低的存储开销和更高的检错率。他们指出关于文件更新及公开验证仍然是未解决的公开问题。他们提出的 HAIL 方案[12]在多个存储服务提供者之间作数据副本或冗余，然后使用 POR 方案检测数据是否被破坏。当检测到某一服务提供者的数据被破坏时，可以利用其他服务器的数据进行恢复。

此外，Yun 等在分析传统 Merkle hash tree 的基础上，提出一种基于 Nonce 的 MAC Tree 方案，将文件块加密后组织为树形结构，以保证数据的保密性和完整性[13]。Wang 等提出一种基于 BLS 同态签名和 RS 纠错码的方法[14]。这些方案也都采用对称加密体制，但是这些技术处理对象的规模有限，且仍然没有考虑数据的动态更新问题，在处理海量云数据时会带来严重的效率问题。

3.2　基于非对称加密体制的方案

几乎与 POR 同时，Ateniese 等人提出了可证明数据持有 (provable data possession, PDP)模型[15]。该模型的主要原理是：令 N 为一个 RSA 模数，F 为代表文件的大整数，检查者保存 $k=F \bmod \varphi N$；在挑战中，检查者发送 Z_N 中的随机元素 g，服务器返回 $s=g^F \bmod N$；检查者验证是否存在 $g^k \bmod N=s$，从而确定原始文件是否存在。原始的 PDP 技术也只能处理静态数据，且计算开销较大，效率不高。

在原始 PDP 技术的基础上，Ateniese 等人提出 SPDP 技术[16]，试图解决原始 PDP 技术效率低下的问题；Erway 等人也提出了 DPDP 技术[17]，试图提供对数据动态更新的支持；此外，Yan Zhu 等结合混合云背景，提出 EPDP 技术以解决多个服务器上多文件的高效持有性证明问题[18]。在原 PDP 技术的基础上衍生了一系列相关的持有性证明技术。综合起来看，这些技术都是基于类似于 RSA 的非对称加密体制。

值得一提的是，Curtmola 等人在 PDP 的基础上创新性地提出针对数据副本持有性验证的 MR-PDP 技术[19]。该技术允许用户通过“挑战−应答”式协议验证云存储服务器真正存储了原始文件的 n 个不同副本，可有效防范服务器同谋攻击。MR-PDP 技术是原始 PDP[15]的扩展，也是基于 RSA 加密机制。它首先将数据加密，然后将加密数据与 n 个不同的随机掩码异或形成不同的数据文件，但是该技术支持新副本的生成，且对 n 个副本的证明开销远小于 n 个不同文件的证明开销之和，具有简单、高效的特点。但是该技术仍然不支持数据的动态更新。

3.3　基于第三方审计的方案

除了以上基于密码学加密技术的方案外，还有一种思想是基于第三方审计的方案。其中比较著名的是由惠普公司 Shah 等提出的一种基于数据委托的审计方案[20] 。该方案将用户的检查任务交给可信第三方来完成，第三方审计者通过“挑战−应答”的方式，通过加密文件的 MAC 值验证存储服务提

供者是否真实地持有一个加密的文件。

第三方审计方案支持公开审计，由可信的第三方代替用户行使验证数据是否真实存在的职能，可在很大程度上减轻用户的负担。但是这种方式只能用于加密的文件，并且要求审计者维护长期的状态信息。一方面，这种技术势必会增加原存储服务商的代价，另一方面，与本文所提出的模型相比，这种技术新增加了一个第三方审计，可能会带来新的安全隐患。

比较而言，第三种方案更关注如何建立可信、可靠的第三方审计，与前两种方案的区别较大，不存在可比性，因此在第 4 节中不将其纳入比较范围之内。

4　方案综合比较

针对第 3 节中介绍的几种主要的数据持有性证明技术，下面根据衡量指标给出一个深入、全面的分析，如表 1 所示。

从表 1 中可以看出：POR、PDP、SPDP 的计算开销和传输开销均较小，为$\Theta(1)$量级，而 HAIL 和 DPDP 则需要$\Theta(\log n)$量级；MR-PDP 因为需要产生 j 个不同的副本，因此需要的计算开销和传输开销也相应较多，但整体仍处于同一量级。在存储开销上，各种方案将文件分块后，产生的上传文件差别并不大。实际上 POR 和 HAIL 方案因为加入了冗余信息导致文件要稍微大一些，但是理论分析表明也不会超过 10%。在数据更新操作的支持上，SPDP 和 DPDP 技术支持相对较好，其中 DPDP 更是支持所有的更新操作。在验证次数上，只有 POR 的表现稍差，其验证次数与进行预处理时插入的岗哨个数有关，数量有限。但是只有 POR 和 HAIL 支持数据的可恢复性，因为只有它们在分块数据中加入了冗余信息，可以容忍一定的数据损毁。而其他方案只能够检测数据是否存在，以及是否完整。如果数据的任何一部分发生了变化，都没有办法恢复出原始数据。

总体来看，POR 通过更新支持和验证次数上的损失而得到可恢复性的支持；而 PDP 在各种开销均较小的情况下支持无限次数的验证，且支持最简单的追加更新操作。这 2 种方案也是最有代表性的方案。其他几种方案针对具体的目标，例如 SPDP 在验证开销上、DPDP 在更新操作支持上、MR-PDP 在对多副本的支持方面取得一定的技术优势。

5　结束语

根据对上述几种方案的分析和比较，笔者认为现有方案存在如下一些缺陷：1) 大部分方案基于公钥密码技术，通常需要进行大量的模幂运算，所以计算开销很大，特别是数据量大的时候效率不高；2) 大部分方案没有考虑数据更新问题，只能用于静态数据的归档存储，不能支持数据的动态操作；3) 大部分方案没有考虑数据恢复技术，虽然可以高置信地检测到数据的损坏，但是却不能正确恢复出原始数据，因此实用性不强。

在对现有方案总结和比较的基础上，笔者认为云存储技术中的数据持有性证明问题仍然有进一步展开深入研究的必要。可能的研究方向包括如下几个方面。

1) 随着数据量的快速增长，用户需要的保存的数据容量将不断膨胀。在此背景下，目前效率较低的各种方案的弱点将逐渐被放大，因此设计具有通信开销小和计算复杂度低的持有性证明方案将是

表 1　　**几种主要方案的衡量指标对比**

方案	文献	计算开销	传输开销	存储开销	更新操作	挑战次数	可恢复性
POR	文献[10]	$\Theta(1)$	$\Theta(1)$	$\Theta(n)$	No	有限	支持
HAIL	文献[12]	$\Theta(\log n)$	$\Theta(1)$	$\Theta(n)$	No	无限	支持
PDP	文献[15]	$\Theta(1)$	$\Theta(1)$	$\Theta(n)$	A	无限	No
SPDP	文献[16]	$\Theta(1)$	$\Theta(1)$	$\Theta(n)$	AMD	无限	No
DPDP	文献[17]	$\Theta(\log n)$	$\Theta(\log n)$	$\Theta(n)$	AMDI	无限	No
MR-PDP	文献[19]	$j\Theta(1)$	$j\Theta(1)$	$j\Theta(n)$	No	无限	No

注：n 表示文件分块的块数；f 为文件上传后被破坏的块与 n 的比值；k 为数据拥有者请求验证的数据块的块数；j 为多副本情况下副本的数量；更新操作中 A 表示 Append、M 表示 Modify、D 表示 Delete、I 表示 Insert；No 表示不支持该功能。

接下来的重点研究方向。

2) 对现有数据修改的需求将会有所增长，包括数据的修改、插入、删除、追加等操作。这就要求研究并支持云存储中动态数据的持有性证明方案。若更新的只是小部分数据，将所有的元数据全部重新计算显然是效率最为低下的一种做法。可能的解决方法是研究支持合成操作的证据，通过在原有证据的基础上叠加修改内容所对应的证据从而形成新的证据。此外，将若干证据合成形成新的证据还可以增加挑战验证的次数，可有效节约计算和存储开销。

3) 随着存储系统规模的不断扩大，部分节点出现故障的可能性也随之增加。除了存储系统硬件本身需要进行健壮性方面的设计外，软件方面也必须提供一定的容错、纠错能力。可能的研究方向是研究带纠错功能的数据持有性证明方案，将数据的持有性验证与纠错编码技术相结合，设计具有纠错功能的冗余存储方案，在发现数据完整性遭到破坏的情况下及时恢复出原始数据。

4) 为了进一步提高数据服务质量保证，在云存储中还需要提供多副本服务。那么，如何验证存储服务提供商确实按照与客户的约定忠实地维持足够的副本数量也将成为研究的一个热点问题，即多副本的持有性验证问题。该问题与数据的持有性问题有一定的联系，但是更具有挑战性，因为设计方案时必须解决服务器之间的“同谋”攻击。

5) 目前海量存储建设的方向是综合了数据处理能力与数据存储能力的数据中心，使用同态加密算法在内容保密前提下提供一定的数据处理能力将是一个热点研究问题。而基于同态标签的验证方案将是非常有价值的一个研究方向。

参考文献：

[1] 冯登国, 张敏, 张妍. 云计算安全研究[J]. 软件学报, 2011,22(1): 71-83.
FENG D G, ZHANG M, ZHANG Y. Study on cloud computing security[J]. Journal of Software, 2011,22(1):71-83.

[2] FOSTER I, KESSELMAN C. The Grid 2: Blueprint for a New Computing Infrastructure[M]. San Francisco: Morgan Kaufmann Publishers Inc, 2003.

[3] 陈康,郑纬民.云计算：系统实例与研究现状[J]. 软件学报,2009,20(5): 1337-1348.
CHEN K, ZHENG W M. Cloud computing: system instances and current research[J]. Journal of Software, 2009, 20(5): 1337-1348.

[4] SENY K, KRISTIN L. Cryptographic cloud storage[A]. FC′OS, Financial Cryptography and Data Security[C]. Roseau, Commonwealth of Dominica, 2010. 136-149.

[5] Amazon. Amazon simple storage service[EB/OL]. https://s3. amazonaws .com/.2012.

[6] Cloud Security Alliance. Security guidance for critical areas of focus in cloud computing[EB/OL]. http://www.cloudsecurityalliance.org/csaguide.pdf.2009.

[7] YUMEREFENDI A R, CHASE J S. Strong accountability for network storage[J]. ACM Transactions on Storage, 2007, 3(3): 6-16.

[8] KUBIATOWICZ J, BINDEL D, CHEN Y. Oceanstore: an architecture for global scale persistent storage[A]. International Conference on Architectural Support for Programming Languages and Operating Systems[C]. Cambridge, MA, USA, 2000. 190-201.

[9] MAHESHWARI U, VINGRALEK R, SHAPIRO W. How to build a trusted database system on untrusted storage[A]. Conference on Symposium on Operating System Design and Implementation (OSDI 00) [C]. San Diego, California, USA, 2000. 10-20.

[10] ARI J, BURTON K. PORs: proofs of retrievability for large files[A]. 14th ACM Conference on Computer and Communications Security[C]. Alexandria, VA, USA, 2007. 584-597.

[11] BOWERS K D, JUELS A, OPREA A. Proofs of retrievability: theory and implementation[A]. ACM Cloud Computing Security Workshop at CCS[C]. 2009. 43-54.

[12] BOWERS K D, JUELS A, OPREA A. HAIL: A high-availability and integrity layer for cloud storage[A]. 16th ACM Conference on Computer and Communications Security[C]. 2009. 187-198.

[13] YUN A, SHI C, KIM Y. On protecting integrity and confidentiality of cryptographic file system for outsourced storage[A]. ACM Cloud Computing Security Workshop at CCS[C]. 2009. 67-76.

[14] WANG Q, WANG C, LI J. Enabling public verifiability and data dynamics for storage security in cloud computing[J]. LNCS 5789. Springer-Verlag, 2009. 355-370.

[15] ATENIESE G, BURNS R, CURTMOLA R. Remote data checking using provable data possession[J]. ACM Transactions on Information and System Security, 2011, 14(1): 12-34.

[16] ATENIESE G, PIETRO R D, MANCINI L V. Scalable and efficient provable data procession[A]. 4th International Conference on Security and Privacy in Communication Networks[C]. Istanbul, Turkey, 2008. 1-10.

[17] CHRIS E, ALPTEKIN K, PAPAMANTHOU C. Dynamic provable data procession[M]. ePrint Archieve, 2009.

[18] ZHU Y， WANG H， HU Z． Efficient provable data possession for hybrid clouds[A]. 17th ACM Conference on Computer and Communications Security (CCS'10) [C]． Chicago, IL, USA, 2010. 756-758.

[19] CURTMOLA R, KHAN O, BURNS R. MR-PDP: Multiple replica provable data possession[A]. 28th IEEE International Conference on Distributed Computing Systems[C]. Beijing, China, 2008. 411-420.

[20] SHAH A, BAKER M, MOGUL C. Auditing to keep online storage services honest[A].11th Workshop on Hot Topics in Operating Systems[C]. San Diego, California, USA. 2005. 1-6.

作者简介：

付伟（1978-），男，湖北武汉人，博士，海军工程大学讲师，主要研究方向为分布式计算、云计算与云安全。

叶清（1978-），男，湖北蕲春人，博士，海军工程大学副教授，主要研究方向为信息安全、入侵检测。

陈泽茂（1975-），男，福建福州人，博士，海军工程大学副教授，主要研究方向为信息安全、无线网络安全。

吴晓平（1951-），男，山西新绛人，博士，海军工程大学教授、博士生导师，主要研究方向为信息安全、应用数学。