|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 分类号 |  | |  | | | 密 级 | |  |
| U D C |  | |  | | | 编 号 | |  |
|  | | | | | | | | |
| 武汉大学logo  **海 量 存 储 技 术 论 文** | | | | | | | | |
|  | | | | | | | | |
| 浅述云存储系统可靠性技术 | | | | | | | | |
|  | | 研究生姓名 | | ： | 储玮 | |  | |
| 学号 | | ： | 2017202110103 | |
| 专业名称 | | ： | 软件工程 | |
|  | |  |  | |
|  | | | | | | | | |
| 二〇一七年十一月 | | | | | | | | |

**摘 要**

云存储系统现在已经足够成熟，可以处理大量的异构和快速变化的大数据。然而，由于云存储系统由大规模硬件组件组成，失败是不可避免的。提高云存储系统中的容错能力对于大型数据应用来说是一个巨大的挑战。复制和擦除编码是云存储系统中使用的最重要的数据可靠性技术。本文探讨了在云存储系统中使用上述两种技术以及大数据应用云存储可靠性面临的挑战。

**关键词：数据可信度；云存储；复制；擦除编码**

**Abstract**

Cloud storage systems are now mature enough to handle large numbers of heterogeneous and rapidly changing big data. However, because cloud storage systems consist of large-scale hardware components, failures are inevitable. Improving fault tolerance in cloud storage systems is a huge challenge for large data applications. Copy and erase coding is the most important data reliability technology used in cloud storage systems. This article explores the challenges of using both of these technologies in cloud storage systems and the reliability of cloud storage for big data applications.

**Keywords:Data reliability; Cloud storage;Replication; Erasure coding**

**目 录**

**中文摘要0**

**英文摘要0**

**1.绪论1**

**2.背景介绍2**

2.1.云存储和大数据应用程序2

2.2.数据失败2

2.3.数据可靠性3

**3.擦除编码5**

**4.复制6**

4.1.静态复制6

4.2.动态复制6

**5.大数据应用云存储可靠性7**

5.1.研究挑战7

5.1.1.存储效率7

5.1.2.带宽效率8

5.1.3.能源效率8

5.1.4.数据访问延迟8

5.3.未来研究方向8

**6.总结10**

**参考文献11**

**1.绪论**

在当代大数据社会中，数据量增长速度超过存储容量。根据国际数据公司（IDC）第六次年度研究报告，到2020年，数字数据将每两年翻一番。云计算提供了一种具有成本效益的方式来支持大数据和分析应用程序，从而推动商业价值。到2020年，数字世界中大约40％的数据将被云计算提供商存储或处理。云存储为大数据的存储提供了合理的可扩展性，有助于处理数据种类，数量和速度的稳定增长。

由于云存储建立在商用服务器和磁盘驱动器上，因此存在可能危及依赖于它的应用程序运行的故障。因为故障是云存储系统的常态，所以提高数据可靠性同时在数据恢复过程中保持系统性能，是在云计算上部署大数据应用程序中最重要的挑战之一。本文介绍了云计算中使用的复制和删除编码两种技术，在提高大数据应用程序的数据可靠性方面的现状和挑战。

本文的其余部分安排如下。在第2节中，我们讨论了云存储中用于大数据应用程序数据故障和数据可靠性的存储系统和文件系统的类型。在第3节和第4节中，我们分别讨论擦除编码和复制中涉及的技术状态。在第5节中，我们介绍了大数据应用的云存储可靠性技术现状。最后，在第6节中，我们总结并提供了这个研究领域未来发展的基础。

1. **背景介绍**

本节将简要讨论用于大数据应用程序的云存储系统中使用的存储系统和文件系统的类型。随后介绍了数据失效和数据可靠性的分析。  
**2.1.云存储和大数据应用程序**  
 云存储系统由许多通过网络连接的存储设备组成，具有存储虚拟化的特点。存储虚拟化是从应用程序抽象物理存储并将逻辑存储映射到物理存储的技术。存储设备网络可以被视为一个单一的存储设备，用户可以访问信息，无论物理位置和存储模式。  
 根据客户端如何访问和接口数据，云存储系统可以分为文件存储，块存储和对象存储。

文件存储：在文件存储中，文件按层次结构组织。有关文件的信息作为元数据存储在存储系统中。可以通过指定单个文件的路径来访问这些文件。它为应用程序提供了更高级别的存储抽象，并支持在不同平台之间传输安全的数据。如果文件和元数据数量有限，则在局域网（LAN）中实现良好的性能。文件服务器维护元数据和授权I / O在多个客户端之间共享文件。但是，文件服务器争用会影响数据检索性能。  
 块存储：在块存储中，文件被分成块，并为每个块分配一个地址。应用程序可以访问和组合块的地址。存储应用程序保留元数据并使用它来共享数据。它没有任何文件服务器来授权I / O，客户端可以使用元数据直接访问存储设备。它具有良好的性能。但是，它没有提供有前途的安全数据传输。  
 对象存储：在对象存储中，文件和元数据被封装为一个对象，并且对象被分配一个对象ID。该对象可以是任何类型和地理分布的。可以为每个对象分配唯一的元数据，例如关联的应用程序对象类型，保护级别，复制次数和地理位置。它从应用程序到存储设备的存储管理。这使得使用元数据的安全的直接数据访问客户端。它提供了出色的可伸缩性来支持大数据应用程序。对象存储除了复制之外，还支持有效的删除编码技术。  
 云存储系统的分布式，虚拟化和可扩展的特性可以很好地适用于大数据的多样性，体积和速度特性。

**2.2.数据失败**  
 在云存储系统中，很多因素都可能导致数据失败。数据故障也导致云服务失败。云数据故障的主要原因是硬件，软件，网络和电源故障。磁盘是基于云存储的核心元素，是最常见的故障组件。Vishwanath和Nagappan[1]分析了大型云计算基础设施的硬件可靠性。如图1所示（2010年Vishwanath和Nagappan收集的数据），78％的服务器故障是由于硬盘造成的，5％是由于廉价磁盘阵列（RAID）控制器造成的，5％是由于内存造成的，其余14％由于其他因素。硬盘是最常被更换的组件，它们是服务器故障的最常见原因。  
 数据失败可能是暂时的或永久的。由于网络中断，节点/机器故障，停电和自动修复过程而导致的数据不可用性是短暂的，不会导致永久的数据丢失。由于硬盘故障或数据损坏导致的数据不可用导致数据永久丢失。  
 云存储系统中的各种组件故障导致永久和暂时的数据故障。磁盘是云存储系统中最重要的组件。如果处理不当，磁盘故障会导致永久数据丢失。云存储系统中的大多数其他组件故障只会导致暂时中断。有些停电可能持续数小时，造成巨大的经济损失。

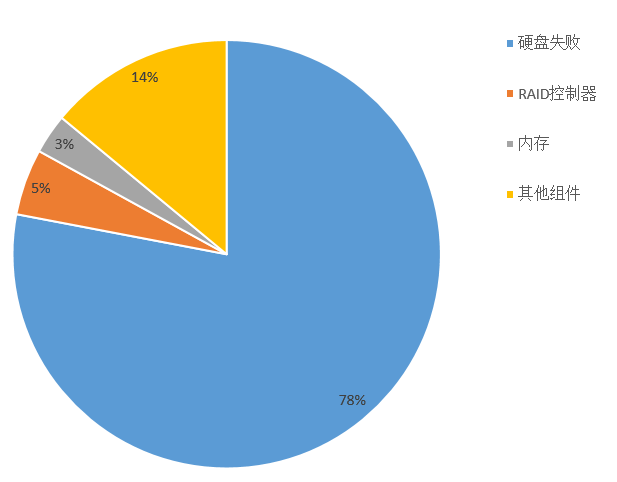


图1. 云计算系统中服务器故障的原因

**2.3.数据可靠性**

数据可靠性包括最大化数据的耐用性和可用性。耐久性减轻了永久性故障，可用性减轻了瞬间故障。  
 如图2所示，云数据中心采用了各种机制来提高存储系统的容错能力。 RAID阵列，可交换驱动程序和纠错码存储器（ECC RAM）可以缓解硬件故障的影响。 RAID阵列是由多个磁盘组成的逻辑单元，用于存储带有条带，镜像和奇偶校验的数据。可交换驱动程序允许管理员在系统保持运行模式时交换失败或预测失败​​的驱动器。ECC RAM用于通过将奇偶校验位与每个二进制代码关联来检测和纠正单个位错误。网络故障和停电分别采用网络冗余和双电源供电。

由于任何问题（包括云存储灾难）的故障都通过删除编码，复制和弹性分布式数据集（RDD）进行处理。复制和删除编码用于处理主要数据失败。RDD用于保护大数据应用程序生成的中间数据。  
 删除编码和复制是云存储中最流行的可靠性机制。在复制中，数据文件/对象被分成块并在存储系统上多次存储。如果请求的数据在一个磁盘中不可用，则从下一个可用磁盘提供[2]。在擦除编码中，数据文件/对象被分成块。奇偶校验数据与原始数据一起被创建和存储，使得如果所请求的数据不可用，则在奇偶校验数据的帮助下重建和服务奇偶校验数据。  
 尽管云提供商利用各种可靠性技术来提高对各种组件故障的容错能力，但复制和删除编码却因地理上分散的冗余而与众不同。因此，复制和删除编码支持包括灾难在内的任何类型的数据丢失。

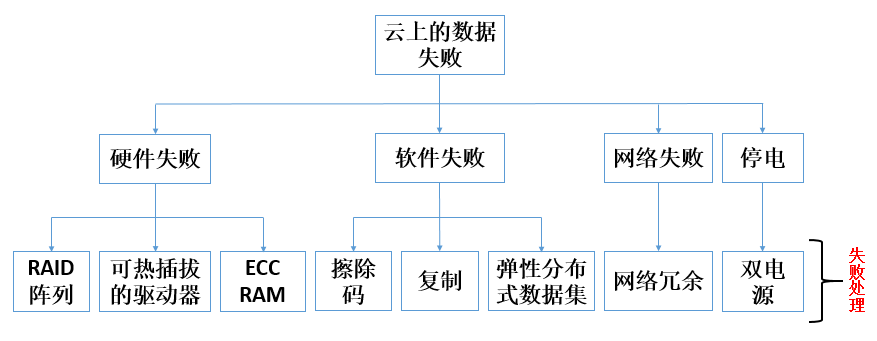
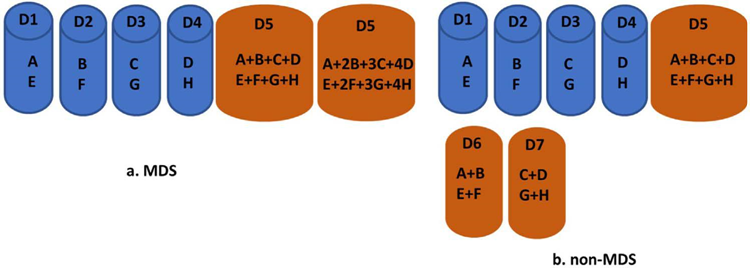


图2. 云数据中心的故障处理

**3.擦除代码**  
 擦除编码在保护大规模存储系统中的数据免于失败方面发挥着主要作用[2]。在云计算出现之前，使用纠删码来检测和纠正存储和通信系统中的错误[3]。在（n，k）个纠删码存储系统中，一个大小为B的文件将被分成k个相等的块，并且添加n-k个奇偶块，使得n个块中的任何k个可以恢复原始文件。用于计算奇偶校验数据的算法可以是标准算术或伽罗瓦域算术[2]。在标准算术中，加法是作为二进制“异或”进行的，而乘法则是“二进制与”。如果一个字中的位数是1，则执行标准算术。当一个字中的位数增加时，使用伽罗瓦域算法计算奇偶性。在Galois域GF（2n）中，算术运算被绑定在从0到2n-1的有限数字集内;加法是按位XOR，乘法更复杂，这取决于硬件，存储器和字中的位数[2]。  
 擦除码可以分为最大距离可分（MDS）和非MDS。如果m个磁盘保存奇偶校验数据并且系统容忍m个磁盘故障的任意组合，则该代码被称为MDS;如果m个磁盘专用于保存校验数据，则非MDS代码只能容忍少数m个磁盘故障组合。例如，在图3a中，磁盘D5和D6专用于奇偶校验，因此该系统可以容忍任何两个磁盘故障。 这使得它的MDS代码。 在图3b中，D5，D6和D7专用于奇偶校验，但不能容忍任何三个磁盘故障。 例如，如果D1，D5和D6同时失效，则D1中的数据将不会被恢复。 这就是所谓的非MDS代码。

简单的擦除代码的例子是RAID-6，阵列代码和Reed-Solomon代码[2]。 RAID-6代码是MDS，它为数据块创建两个奇偶校验块，以便处理两个磁盘故障。数组代码是用标准算术（即XOR运算）实现的。在数组码中，奇偶性是以系统数据（原始数据）的不同线性组合来计算的。行对角奇偶校验（RDP），EVENODD，Blaum-Roth和解放代码是RAID 6的阵列代码，可以容忍最多任何两个磁盘故障。星码是一个数组码，它可以容忍三个磁盘故障的任意组合。 Cauchy Reed-Solomon，广义EVENODD和广义RDP是可以为k和m的任意值定义的数组代码[2]。最近的进步减少了Reed-Solomon码的Galois域算法的CPU负担。此外，为k和m的任何值定义Reed-Solomon代码是很简单的。



翻译时出错图3.不同的删除编码类型a.MDS代码b.非MDS代码

**4.复制**  
 复制是云数据中心最常用的可靠性机制，以便以低延迟和最小的带宽消耗来提高可用性和持久性。失败时，为了保持持久性，需要在活动磁盘中恢复失败的副本。这种恢复可以被动地或主动地执行。在反应式复制中，副本将在失败后创建。在主动复制中，即使在发生故障之前也会创建副本。用于复制的常用方法是静态和动态复制。  
**4.1.静态复制**  
 在静态复制中，副本的数量及其位置是固定的。无论用户行为如何变化，副本都是手动创建和管理的。随机复制是HDFS，RAMCloud，GFS和Windows Azure中最常用的复制技术。在这种技术中，数据被复制到不同机架上随机选择的节点上。随机块复制可以容忍并发失败，因为块被放置在不同的机架上。但是，当所有副本都丢失时，它是无效的。另外，修复丢失的数据涉及与定位和恢复丢失的数据相关的高成本。  
 Liu和Shen[4]提出了多重故障弹性复制（MRR）来提高云存储的可用性。作者根据对象的流行度为每个对象定义了不同数量的复制。节点被分成不同的组，使得组可以处理不同数量的复制，每个节点由来自不同数据中心的节点组成。它降低了一致性维护成本低的数据丢失的可能性。 Long等人[5]提出了云存储的多目标优化管理（MOM）算法。 MOM根据具有五个目标的数学模型（即不可用性，服务时间，负载变化，能量消耗和延迟）来决定副本的数量和副本的位置。在模型的定义中考虑了参数大小，文件访问速率，故障概率，传输速率和容量数据节点。作者表示，该算法增加了文件的可用性，负载平衡，减少了服务时间，延迟和能源消耗。

**4.2.动态复制**

在动态复制中，副本是动态创建和删除的。为了提高耐用性，可用性，成本，存储效率，带宽，延迟，能源和执行时间，根据用户行为自动处理副本的创建，位置，管理和删除。Boru等人[6]提出了一种数据复制技术来优化云数据中心的能耗，网络带宽和通信延迟。他们定义了能耗和带宽需求的模型，并提出了一个基于此模型的能量有效复制策略，可以减少通信延迟。

**5.大数据应用云存储可靠性的最新技术**

由于云存储系统的故障频繁，云存储系统采用数据冗余来处理故障。复制是提高数据可靠性的简单解决方案。但是复制字节和PB数据会大大增加存储开销。现在擦除编码较受欢迎，因为它在存储方面节省了大量的可靠性和耐用性。但是，恢复丢失的数据所涉及的重建成本平衡了存储节省。与复制相比，Reed-Solomon代码每比特需要大约十倍的修复开销。本节的其余部分将讨论在云存储系统中使用大数据应用冗余技术所面临的挑战。  
 一些研究侧重于减少网络流量，减少擦除编码存储系统中重建失败数据的磁盘I / O。很少有作品提供额外的存储开销来提高擦除编码存储系统的性能，但是没有一个可以使擦除代码的性能像大数据应用程序的复制所达到的性能一样。  
 一些研究侧重于最小化复制存储系统中的冗余数量，以提高存储效率。与纠删编码相比，没有任何一个可以显着减少存储开销而不会牺牲可靠性。实现可靠性，存储效率和性能以及复制或擦除编码的存储系统尚未实现。

混合可靠性机制可能是未来数据中心的选择。混合可靠性机制结合了复制和时代编码。双重编码是在一个对等体中保留一个完整的数据副本，并且擦除在网络中传播的编码片段。在双重编码中，原始数据片段和奇偶校验片段的副本被安排在网络中的不同对等体中。即使在重建时节省带宽，也会影响存储效率。李等人[7]提出了主动纠删编码（ProCode），它基于驱动器故障预测自动调整数据的复制因子。它将读取延迟降低了63％，重建时间缩短了78％。此ProCode在由闪存驱动器组成的存储系统中没有任何作用，可交换的驱动程序可以更有效地处理驱动器故障。

**5.1.研究挑战**  
 云存储系统中大数据应用的数据可靠性领域面临着许多开放的挑战。在本节中，我们将研究提高云存储系统中大数据应用程序的数据可靠性所面临的挑战。

**5.1.1.存储效率**  
 复制的数据可靠性直接与存储开销成正比。因此，在不牺牲可靠性的情况下提高存储开销是复制中最大的挑战。尽管删除编码可以大大节省存储空间并且具有相当的可靠性，但是在出现故障的情况下，它会增加大数据应用程序的网络流量和延迟。本次调查针对复制方面的各种研究，尝试使用动态复制策略来节省存储空间，但是却提高了可靠性。这也表明，纠删码的各种研究集中在减少网络流量和延迟。实现动态冗余的自动化，比如改变擦除编码数据中的副本数量以支持故障和使用率峰值，还可以进一步提高存储效率。

**5.1.2.带宽效率**  
 网络带宽始终是分布式存储中的稀缺资源。带宽使用量与分布式存储中传输的数据量成正比。在复制和擦除编码的存储系统中，修复失败的数据消耗大量的网络带宽。传统的擦除代码比复制涉及更多的带宽消耗。节点故障是触发云数据中心数据重建的事件，增加了纠删码存储系统中的网络流量。在擦除编码存储系统中减少网络流量的各种工作已经被突出强调。但是，这些不能减少复制存储系统的网络流量。即使最近改进了纠删码，涉及重构的网络传输也是最重要的挑战之一。在不牺牲性能和能源效率的情况下提高带宽效率是最重要的挑战之一。应该改进故障预测技术以发现云存储系统中的数据故障。即使在故障发生之前，故障预测数据的动态复制也可以显着减少擦除编码存储器中的重建次数，从而减少网络流量。

**5.1.3.能源效率**

存储系统是云计算中最重要的耗能组件之一。数据中心使用的能源效率方法节省了运营成本，并有助于保护环境。  
 在纠删码中存储的节能专用于重构数据，反之亦然。因此，在不牺牲存储效率的情况下改进数据重建是一个挑战。通过减少纠删码中的重构次数，可以减少网络流量，提高能源效率，通过对所选数据块进行季节性复制，可以减少重建次数。

**5.1.4.数据访问延迟**  
 在擦除编码的存储系统中，当出现故障时，必须执行解码以重建原始数据。因此，数据访问延迟是擦除编码存储系统中最重要的挑战之一。在复制存储系统中，通过选择最佳的复制位置可以显着降低访问延迟。大数据应用程序的能耗和性能可以通过降低访问等待时间来考虑改进。减少存储开销减少延迟是研究人员的一个挑战。通过基于擦除代码中的访问历史记录和失败日志激活主动动态复制，可以减少存储开销，从而减少访问延迟。

**5.2.未来的研究方向** 在大数据时代，提高对数据失效的容错带来了各种挑战，如如何减少存储开销；如何减少网络带宽和/或网络流量;如何最小化延迟最后如何最大限度地减少能源消耗。为了解决这些问题，将来的研究方向如下：  
 （1）在复制中减少存储开销的新技术，不会影响数据的持久性和可用性：应该开发应用主动式故障处理和动态复制的技术，以降低复制的存储开销，而不会牺牲大数据应用程序的可靠性。  
 （2）大数据优化的删除编码：删除编码有助于减少存储开销，同时保持持久性和可用性。但是，这是以增加网络流量，高磁盘I / O和大延迟为代价的。因此，需要一个新的删除编码来减少网络流量，磁盘I / O和延迟，同时减少大数据的存储开销。  
 （3）提高数据可靠性的混合技术：由于复制和纠删编码都有优势和局限性，因此利用混合技术提高可靠性可以充分利用每种技术的最佳方面，这是一个很有希望的研究课题。  
 （4）基于故障预测的片段动态复制：虽然擦除编码可以降低存储系统的能耗，但这是由故障发生时重建数据所需的额外能量所决定的。因此，如果算法可以预测由于故障而更可能变得不可用的项目，那么如果预测项目被主动复制，则可以节省能量，因为在这种情况下可以避免重构。

另外，还需要进一步考虑根据用户可靠性要求优化复制和纠删编码配置参数的方法以及为大数据应用程序定义数据可靠性时云存储的地理多样性。

**6.总结**

云计算在服务大数据应用方面扮演着主要角色，因为它提供了具有成本效益的按需服务。云计算使得存储和计算资源可以根据需求快速升级和降低。随着云存储系统中的故障逐渐成为常态，云存储系统采用了各种容错机制来提高数据的可靠性。由于复制涉及巨大的存储开销，擦除编码在云存储系统中获得了大数据应用的推崇。但是，如果发生故障，重建丢失的数据会涉及许多影响应用程序性能的资源。这可以防止云存储系统转向擦除编码。在本文中，我们讨论了这两种技术的现状和面临的挑战，并总结了未来的研究方向。

**参考文献**

[1] Vishwanath K V, Nagappan N. Characterizing cloud computing hardware reliability[C]// ACM Symposium on Cloud Computing, SOCC 2010, Indianapolis, Indiana, Usa, June. DBLP, 2010:193-204.

[2] Plank J S. Erasure codes for storage systems : a brief primer[J]. ;login:: the magazine of USENIX & SAGE, 2013, 38:págs. 44-50.

[3] Vins V A, Umamageswari S, Saranya P, et al. A survey on regenerating codes[J]. 2014, 4(11):2250-3153.

[4] Liu J, Shen H. A Low-Cost Multi-failure Resilient Replication Scheme for High Data Availability in Cloud Storage[C]// IEEE, International Conference on High PERFORMANCE Computing. IEEE, 2017.

[5] Long S Q, Zhao Y L, Chen W. MORM: A Multi-objective Optimized Replication Management strategy for cloud storage cluster[J]. Journal of Systems Architecture, 2014, 60(2):234-244.

[6] Boru D, Kliazovich D, Granelli F, et al. Models for efficient data replication in cloud computing datacenters[C]// IEEE International Conference on Communications. IEEE, 2015:385-402.

[7] Li P, Li J, Stones R J, et al. ProCode: A Proactive Erasure Coding Scheme for Cloud Storage Systems[C]// Reliable Distributed Systems. IEEE, 2016:219-228.