重复数据删除技术研究报告

班级：硕士五班

学号：2017282110215

姓名：刘勇琰

目录

摘要：

在存储系统中自动消除重复数据，通常被称为重复数据删除，旨在减少存储系统中使用的存储容量，被认为是一种降低存储成本的有效技术。因此，它已经被应用到不同的存储类型，包括归档和备份、主存储、固态存储，甚至是随机访问内存中。尽管所有存储类型都共享重复数据删除的一般方法，但每种存储类型都有特定的挑战。本文简单介绍了重复数据删除系统及其与压缩的关系，然后对数据删除系统面临的挑战做了分析，最后介绍了提高重复数据删除系统性能的几种常用方法和数据删除技术未来的发展趋势。

1 重复数据删除系统介绍

1.1 重复数据删除技术

自动清除重复数据已经在档案和备份系统中使用了很长时间，并且已经成为许多存储系统的特征功能。随着数字存储信息的空前增长，重复数据删除正在受到越来越多的关注。重复数据删除是一种自动消除粗粒度和不相关重复数据的技术。不同与传统的压缩技术消除文件内的冗余或小部分文件间的冗余（通常一起存储在同一操作中），重复数据删除旨在消除大数据集下文件内和文件间的冗余（由用户在不同的时间存储，甚至可能跨越多个分布式存储服务器）。

不同的存储环境具有特定的要求，因此需要设计相应的重复数据删除系统。在某些情况下，数据预计是不可变的，永远不会被删除；此外，延迟通常不如吞吐量重要，因为数据的恢复只是零星的，而大量的数据必须在有限的时间窗口中归档；另一方面，备份中的恢复操作会更加频繁，并且旧的数据可以被删除，因

此需要有效的参考管理和垃圾收集机制。

主存储中活动数据的重复数据删除改变了备份和存档基础设施所做的许多假设。活动数据不再是不变的，对参考管理和垃圾收集机制有深远的影响。另外，使用主存储的应用程序对磁盘 I/O 操作的延迟有严格的要求，这限制了重复数据删除可能在 I/O 关键路径中引入的开销。读取请求非常频繁，并且要比备份和存档存储的恢复操作更有效地服务。

在具有进程分支和共享库的现代操作系统中，已经避免了随机存取内存中的重复页面，但是虚拟化技术没有充分利用这些方法。实际上，具有相同操作系统和共享库的多个虚拟机（VM）在同一主机上运行，也不会共享重复页面。重复数据删除可用于消除这些虚拟机中的冗余页面，然而仅限于在同一台主机上的多个虚拟机间进行重复数据删除。因此，与其他存储环境不同的是，在基础架构中扩展到较大群集并不是问题。相反，内存需要重复数据删除机制来保持严格的 I/O 性能和充分利用节省空间的元数据。

在固态存储器(SSD)中也应用了重复数据删除技术，这不仅增加了这些设备的可用存储空间，还增加了它们的寿命。由于重复的写请求可以在实际存储之前被拦截和共享，所以可以减少写操作到磁盘上，因此，也可以延长存储器的使用寿命。SSD 的重复数据删除系统限制令元数据的内存空间，并且必须对 I/O 性能产生最小影响。

重复数据删除的效率是通过重复数据删除的增益得到，定义为删除的重复副本数量，从而减少保留相同数据所需的存储空间。云计算和虚拟服务器技术的发展，可以将重复数据删除应用于通用的存储系统映像中。

通过重复数据删除节省的存储空间降低了基础设施的成本，还可以通过额外

的 RAID 配置来提高存储可靠性。重复数据删除能够提高整个存储管理堆栈中缓存的性能和 I/O 的效率，同时能够减少网络带宽的消耗。

1.2 重复数据删除系统

重复数据删除可视为同一数据在两种不同视图之间的双向映射：包含可识别重复逻辑视图和删除了重复项后存储在实际设备中的物理视图。映射过程体现在生成和使用数据的应用程序与存储设备之间的 I/O 路径中。图 1 描述了每个视图，并指出了每个视图中导致了不同的设计决策和权衡的关键问题。

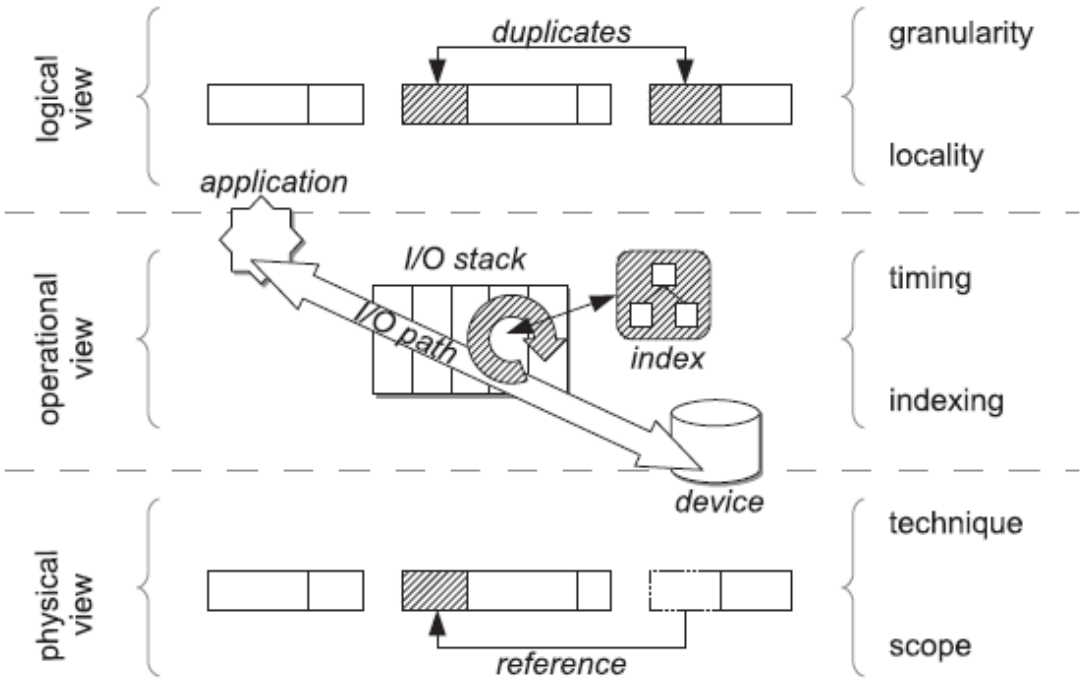

图 1 重复数据删除系统结构

因此，重复数据删除系统中数据的逻辑视图是一组关于工作负载的假设，用于确定哪个重复内容是相关的，存在哪些重复内容以及哪些重复内容应该删除。所有重复数据删除系统将数据划分为要进行比较的离散块，标识其中重复的块，并最终删除。这种分区可以用不同的粒度来完成，在空间或时间上发现重叠块可能性相近的假设导致利用局部性的设计决策影响重复数据删除过程的效率和有

效性。

另一方面，重复数据删除系统中数据的物理视图关注于磁盘上删除重复数据使用的技术，从而实现逻辑视图的高效重构。鉴于分布式存储系统的当前相关性，关键设计决策是重复数据删除技术的分布范围。这可以被定义为能够在不同节点上表示被删除的重复数据，这样就可以对数据的重新构建进行协作。

重复数据删除是存储管理系统中的一个过程。这将 API 暴露给客户端应用程序（如文件系统或块设备），并由多个堆叠的软件层和处理，网络和存储组件组成。这里的关键设计问题是重复数据删除操作的时间，比如搜索重复数据，关于 I/O 操作的关键路径。由于查找重复是潜在的资源密集型操作，因此它总是基于支持重复块的高效匹配的索引数据结构。因此，索引方法不仅对重复数据删除过程的效率有很大的影响，而且还有可能影响速度的准确性。

1.3 重复数据删除与压缩的关系

目前数据缩减(Data Reduction)技术是存储系统领域非常重要的技术，包括重复数据删除技术和数据压缩技术，压缩技术的优点是非常成熟，并且易于理解，它的缺点是处理机制仅限于单个文件之内，而无法做到跨文件处理。压缩通常是针对那些存取频率不是很高的数据，这是因为数据的压缩和解压缩需要 CPU 进行非常密集的计算处理，计算开销较大，这样往往会影响数据的访问。重复数据删除采用指纹技术来处理数据块的内容，有相同 Hash 值的块只存储一次来实现数据的消重，在存储层使用重复数据删除技术并不意味着不再需要数据压缩等其它数据缩减技术。重复数据删除是一种补充型技术，能够与包括业界标准 LZS 压缩算法在内的现有数据压缩技术形成互补，以提供双重数据缩减性能。重删系统

必须先实施重复数据删除而后再压缩，因为压缩会使即使非常相似的内容经压缩后变得非常不同。

2 重复数据删除面临的挑战

2.1 开销和增益

重复数据删除系统面临的主要挑战是，重复数据删除的增益和存储系统的开销之间的权衡。较小的块大小增加了重复数据删除率，能够节省更多的空间，但却导致了更大的索引结构，这些结构的维护成本更高。理想情况下，索引将被完全加载到内存中，但是对于大型存储系统和相对较小的块大小，索引会非常大，必须存储在磁盘上。这增加了重复数据删除所需要的磁盘 I/O 操作，这可能会影响前端的性能。

另外，只要数据进入存储系统就应该执行重复数据删除，以最大化其增益。但是，查找重复项是一项资源密集型任务，如果在存储写入的关键路径中执行，将会影响延迟。如果重复数据删除从关键路径中删除并在后台完成，则需要额外的临时存储，并且必须从存储中读取数据以查找重复项，从而增加了存储 I/O 带宽的消耗。

省略的数据块越多，重复数据删除后数据的物理布局越不同于原来的布局。也就是说，重复数据删除引入了碎片化，这将降低读取和恢复操作的性能。对于正确地重建重复数据，还需要额外的元数据。因此，维护这些元数据的完整性需要额外的开销，因为在修改或删除它之前，必须确保某个共享块不再提供任何的 I/O 操作。更具体地说，这需要管理对共享块的引用和垃圾收集机制，可能会对

性能造成一定影响。

2.2 可扩展性

原则上，任何块可以与任何其他块进行比较，如果找到匹配项，则可以获得最大的增益。但是，这样的完全匹配比较困难，因为大型存储系统中的数据和组件的数量增加了。具体来说，一个集中的索引解决方案很可能会变得非常大，而它的操作是重复数据删除的瓶颈。可以只匹配副本子集的部分索引提高了可伸缩性，但只执行部分重复数据删除。然而，块的数量不能通过探索数据的位置将更相似的块组合在一起来减少。

在分布式存储系统中，一个简单的可扩展性策略是在每个节点中独立执行重复数据删除操作，从而拥有多个独立索引。同样，这种方法只允许部分重复数据删除，因为相同的块可能在多个节点中重复。可以通过将包含匹配数据的更大可能的数据块定向到相同的节点，从而减少重复数据删除的机会。

可伸缩性和增益之间的权衡可以通过使用分布式散列表(DHT)作为被所有节点访问的索引来改进，以一种精确的方式消除了全局的重复。但是，需要对索引进行远程调用，以找到重复或类似的块。如果索引在关键的 I/O 路径中被访问，那么它可能会导致严重的延迟。

2.3 可靠性、安全性和隐私性

分布式的重复数据删除系统必须容忍节点崩溃、数据丢失等故障。消除所有重复的数据也将消除容忍数据丢失和损坏所必需的所有冗余，因此必须维护一定的复制级别。元数据必须具有对故障的适应能力，因此需要长久存储，这也减少

了重复数据删除对空间的节省。另外，数据和元数据必须在大型系统中分布存储，以容忍单节点故障，保持高可用性。

一些重复数据删除系统共享来自不同客户的数据，为了提高了隐私和安全问题，可以通过重复数据删除节省空间来解决。安全性和隐私问题不仅会出现在云存储基础设施中，还会出现在远程存储设备中，存储来自多个客户的数据。

3 重复数据删除系统的性能提高

重复数据删除系统 I/O 路径上分为 Chunking、Chunk Hashing、FP indexing 和 Data Storing 四个阶段，针对其中的每一个阶段既可以采用硬件加速技术又可以采用软件加速技术来提升重复数据删除系统性能。

3.1 FP indexing 查询优化技术

指纹库查询过程的优化是重点，目前针对较大数据量的内存查询优化策略，采用挖掘局部性特征的技术和位置特性(相似性技术)来减少内存的占用和磁盘 I/O 等等；采用 FP 快速判别算法 Bloom filter 技术把常用的热指纹放到内存中，所做的工作都是在重复数据删除率、I／0 吞吐率、响应时间之间进行平衡和折中。

3.1.1 局部性技术

重复数据删除系统随着数据规模的增加，其指纹索引表会逐渐加长，不可能全部放于内存，指纹的查询导致额外的磁盘 I/O，增加了索引查询的时间。为了提高在 RAM 中指纹查询的命中率，降低指纹查询开销，在重复数据删除系统中提出了利用局部性技术(Locality Preserved Caching)来提升指纹的查询优化技术，

局部性技术分为空间局部性(Spatial Locality)和时间局部性(Temporal Locality)。

3.1.2 相似性技术

相似性检测技术在目前的重复数据删除系统中被普遍采用，由于重复数据删除系统是在原有正常 I/O 路径的流程中加入了一个具有重复数据删除功能的模块，该模块的主要功能就是检测输入的新块是否是重复的块。可以引入了相似性技术实现快速的查找比对。在比对数据块的时候只比对满足某一阈值的连续块，这样在重复数据删除的时候只删除连续的数据块，这里存在一个不完全重复数据删除的问题，牺牲了重复数据删除率，但是它能够有效地提高重复数据删除系统的响应时间，这对于主重复数据删除系统是非常重要的。

3.1.3 Bloom Filter 技术

Bloom Filter 是一种空间效率很高的随机数据结构，它利用位数组来很简洁地表示一个集合，并能判断一个元素是否属于这个集合，Bloom Filter 算法的核心思想就是利用多个不同的 Hash 函数来解决“冲突”。在最早的重复数据删除系统 Venti 中采用传统段桶式的 FP index 查询方式，但是当数据规模逐渐增大时，其 FP index 不可能全部放于内存，Bloom Filter 技术是一种利用空间换时间的技术，因此引入 Bloom Filter 技术可以有效提高元数据的查找速度。

3.2 硬件加速技术

从计算机系统结构的观点出发，除了上面基于软件角度的考虑外，重复数据删除系统流程的每一个功能部分都可以采用硬件加速技术来提升重复数据删除系统的性能。

采用多核处理器或多处理器来实现重复数据删除系统的流水线和并行性有助于降低系统延迟时间，该思想的实施将会使主存储重复数据删除系统采用变长分块技术成为可能。SSD 具有较高的 IOPS 数、高吞吐率和低访问延迟的特性，使用 SSD 来存储元数据，来提高元数据的查找速度。而且，在磁盘阵列里可以采用 RAID 技术来进一步提升 I/O 性能。Flash 存储器的读写性能好于 SCSI 磁盘而低于 RAM，使用 Flash 存储器和 Flash 存储器感知的算法和数据结构来获得更快的索引查询操作。但是 Flash 的随机更新是相当低的，通过将 Flash 的随机更新转变为 Log 结构的扩展操作来克服对 Flash 的随机更新。

3.3 垃圾回收技术

垃圾回收技术在存储系统中用于判断某个物理块是否应该被回收，具体来讲就是维护元数据表中的物理块组的使用情况，目前的技术主要有：标记和扫描技术(Mark and Sweep)、基于引用计数(Reference Count based)和基于时效期的技术(Expiry Time based)。

标记和扫描技术(Mark and Sweep)就是对已使用的每一个物理块进行标记，然后对所有的物理块进行扫描以回收未作标记的物理块，但是在标记阶段需要冻结元数据表从而不允许用户访问，另外标记物理块的开销也将是非常大的，不可能将其放于内存中，因此更新 PBA 将会导致低磁盘 I/O 性能。

基于引用计数(Reference Count based)简单来讲就是对每一个物理块使用一个计数器来记录其引用的次数，每一次创建一个新块和删除一个块时，其计数器就相应地加 1 或减 1，每一次更新的时候检测其计数器为 0，说明该块未被使用可以回收，比起标记和扫描的方式来讲避免了耗时的扫描，但系统规模比较大时，

其更新的过程也会降低系统的性能。

基于时效期的技术(Expiry Time based)通过 PBA 时效期的方式来替代引用计数，因而可进一步地减少更新次数，但是当覆写物理块时需要更新其时间，在垃圾回收时需要扫描所有的物理块，看其是否过期。

3.4 striping 技术

重复数据删除造成了数据在磁盘上的不连续存储，从而降低了磁盘 I/O 的性能，Striping 技术可以自动地将 I/O 的负载均衡到多个物理磁盘上的技术，在读数据时以并行的方式从不同的磁盘上获取数据块，从而获得非常好的性能。

4 重复数据删除技术的发展趋势

随着云存储技术、专用处理器技术的进步，加之实时联机数据处理和存储预算压力的考虑，重复数据删除技术正面临从备份/归档存储向主存储环境的转变，数据构成特征缺乏局部性，随机小写比较多，读数据占绝大多数，因此如何提升主存储重复数据删除系统读请求的性能是重复数据删除技术急需解决的问题。

随着存储系统容量的不断扩大，单节点存储系统存在单点性能瓶颈和单点失效的问题，重复数据删除技术面临从单节点向分布式集群环境的转变，如何协调和保障分布式环境下数据完整性、一致性和系统可扩展性仍然有待进一步研究。

随着存储系统数据规模的不断扩大，数据集缺乏局部性特征，其 FP 查询管理开销将逐渐增加，严重地影响了重复数据删除系统的性能，现有的 FP indexing 查询优化技术有待进一步提升，如何通过元数据的布局来提升重复数据删除系统性能将会成为重复数据删除技术研究亟待解决的问题。

针对不同应用场景数据构成特征的复杂性，如何通过结合统计学和数据挖掘领域的各种技术，对数据特性进行充分的分析和挖掘，开发适合不同数据特征的检测技术还有待进一步深入研究。

参考文献：

[1] Paulo J, Pereira J. A survey and classification of storage deduplication systems[J]. ACM Computing Surveys (CSUR), 2014, 47(1): 11.

[2] Shilane P, Chitloor R, Jonnala U K. 99 Deduplication Problems[C]//HotStorage. 2016.

[3] Guo F, Efstathopoulos P. Building a high-performance deduplication system[C]// Usenix Conference on Usenix Technical Conference. 2011:25-25.

[4] Quinlan S, Dorward S. Venti: A New Approach to Archival Storage[C]// Conference on File and Storage Technologies. USENIX Association, 2002:89-101.

[5] Kasavajhala V. Solid State Drive vs. Hard Disk Drive Price and Performance Study[J]. 2011.

[6] Debnath B, Sengupta S, Li J. ChunkStash: speeding up inline storage deduplication using flash memory[C]// Usenix Conference on Usenix Technical Conference. USENIX Association, 2010:16-16.

[7] Xia W, Jiang H, Feng D, et al. Accelerating data deduplication by exploiting pipelining and parallelism with multicore or manycore processors[C]//Proc. 10th USENIX Conference on File and Storage Technologies (FAST). 2012: 14-17.

[8] Data striping[EB/OL]. http://en.wikipedis.org/wiki/Data_striping, 2012-08-

15/2012-08-23.

[9] Lillibridge M, Eshghi K, Bhagwat D, et al. Sparse Indexing: Large Scale, Inline Deduplication Using Sampling and Locality.[C]// Usenix Conference on File and Storage Technologies, February 24-27, 2009, San Francisco, Ca, Usa. Proceedings. DBLP, 2009:111-123.