doi:10.3969/j.issn.1000-436x.2014.10.011

云存储系统中数据副本服务的可靠性保障研究

黄昌勤[1,2]，李源[1]，吴洪艳[1]，汤庸[2]，罗旋[1]
（1. 华南师范大学 教育信息技术学院，广东 广州 510631；2. 中山大学 计算机系，广东 广州 510275）

摘 要：以数据节点与网络链路的可靠性因素分析为基础，提出了云存储系统的数据副本服务可靠性模型。根据访问可靠性与数据副本数量、用户访问量之间的关系，设计数据服务可靠性、副本生成时机、存储节点选择的确定方法，实现了副本分布、删除算法，并在云存储系统 ERS-Cloud 上进行一系列实验，结果表明该方法能够有效保障数据服务的可靠性，进一步降低副本的冗余存储数量。

关键词：云存储；数据副本；可靠性模型；保障
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)10-0089-09

Modeling and maintaining the reliability of data replica service in cloud storage systems

HUANG Chang-qin[1,2], LI Yuan[1], WU Hong-yan[1], TANG Yong[2], LUO Xuan[1]
(1. School of Educational Information and Technology, South China Normal University, Guangzhou 510631, China;
2. Department of Computer Science, Sun Yat-Sen University, Guangzhou 510275, China)

Abstract: The reliability of data-nodes and the reliability of relevant network links were analyzed, and then the reliability model of replica service of cloud storage systems was constructed. According to the relationships among access reliability, the number of replicas and the number of user's accesses, the reliability of data service and the trigger mechanism of replica generation were presented, and the storage node selection was aptly checked, then the replica distribution algorithm and replica deletion algorithm were proposed. Finally a series of experiments were conducted in the cloud storage system, named ERS-Cloud, and the results indicate that the approach can ensure the reliability of data service, and further decrease the number of replicas of the redundant storage.

Key words: cloud storage; data replica; reliability model; maintenance

1 引言

云存储系统的数据存储过程是将资源文件分割成数据块，并根据一定的副本策略分布在不同的数据节点上，以确保数据资源的可靠性，如 GFS[1]、HDFS[2]。由于在副本数量达到一定量的时候，增加副本数量对数据可靠性提升将不再明显[3]，反而会造成存储空间浪费，因此一般云存储系统将副本限额设定为一个不大的值,如默认值 3。在云存储中数据副本的生成与处理也为系统带来了额外的开销，因此副本管理的效果与代价问题也自然成为了许多学者关注的热点。如 Nicolas[4]等提出了一个自我管理、容错和可扩展云存储副本机制，该机制能够根据代价−效率为应用动态分配资源；Liao[5]等提出了一种基于服务质量感应的动态数据副本删除策略，并实现相关算法 DRDS，以降低云存储系统空间占用与

收稿日期：2013-09-18；**修回日期**：2013-12-04
基金项目：国家自然科学基金资助项目(61370229,61370178,61272067,60940033)；广东省自然科学基金资助项目（S2013010015178, 10151063101000046）；广东省科技计划基金资助项目（2012A032200018,2010B010600033）；广东省教育厅科技创新基金资助项目（2012KJCX0037）；中国博士后科学基金资助项目（201003374, 2013M540658）
Foundation Items: The National Natural Science Foundation of China (61370229, 61370178, 61272067, 60940033); The Natural Science Foundation of Guangdong Province (S2013010015178, 10151063101000046); The Science-Technology Projects of Guangdong Province (2012A032200018, 2010B010600033); The Science-Technology Project of DEGP (2012KJCX0037); The Postdoctoral Foundation of China (201003374, 2013M540658)

代价消耗；宋娅菲[6]提出了一种基于竞标模式的副本放置策略以解决云存储系统中副本动态调整的问题，该策略将负载、副本距离等因素转换成竞标参数进行调整。然而，对于绝大多数资源文件，特别是存储时间短、访问量较小的资源文件，仅仅考虑副本冗余也将造成系统资源的大量浪费，同时影响用户访问效能[7]。结合实际应用、数据访问可靠性关注副本管理，也是本领域的关注焦点之一，典型研究有基于用户访问热度进行副本调度，最终优化云存储结构[8,9]；运用副本前测管理副本有效性，从而适当减少副本存在数量[10]；基于可配设置调整云存储容错级别和冗余度，提高云存储的可靠性[11]。后两者与本研究紧密相关，然而，文献[10]主要兼顾存储单元的可靠性并以减少冗余度为目的，文献[11]则从文件的合理分拆与合并角度来兼顾存储容错度。这与本研究中基于数学的建模方法、面向整体的数据服务可靠性视角存在显著差异。

综合上述内容，现有研究多关注副本对云存储系统负载、效率的影响，面向利用率、可靠性的副本管理策略、配置问题等。如何在确保可靠的数据存储和数据应用服务的同时，降低数据副本的冗余程度是当前云存储机制优化亟待解决的主要问题之一。为此，本文将针对云存储设备、网络链路的可靠性建模，并结合数据节点的访问强度，给出判断数据副本服务的可靠性程度的方法，通过研究提出适应于资源云存储的副本服务可靠性保障策略。其中,借鉴领域中的通用可靠概念[12],将数据副本服务可靠性界定为云存储系统中数据副本能被不失效访问的概率。

2 系统相关描述与定义

在云存储系统中，数据副本是各应用的访问数据对象，数据副本宿居于数据节点且由节点的存储设备提供存储服务，并经通信通道完成最终的访问服务。因此，存储设备（含拓扑结构）、存储服务及网络链路等与数据副本服务紧密相关。为了方便数据副本可靠性建模，对存储设备做如下定义。

定义 1 存储设备

存储设备指对外提供数据块存储服务的存储节点设备 D，用五元组 $D=(C_{\text{cpu}}, M_{\text{com}}, N_{\text{port}}, L_{\text{net}}, v_d)$ 表示。

C_{cpu} 表示存储服务计算的核心部件，即设备 D 的 CPU；

M_{com} 表示设备 D 的数据存储部件，为云存储提供数据存储空间；

N_{port} 表示设备 D 对外网络传输接口，保证设备对外提供存取服务的指令与数据信号发送、接收；

L_{net} 表示设备 D 所在的网络位置，与 N_{port} 一起决定了存储服务设备的网络通信质量；

v_d 表示存储服务设备的内部访问执行速度。

定义 2 存储服务

存储服务是指数据节点对外提供数据块的访问服务，包括对用户提供数据的存储或读取服务；在服务节点间的数据调度过程中，数据节点作为目标存储节点或数据原始节点对其他数据节点提供数据存储或读取服务。由此存储服务 WSS 的任务可表示为一个三元组 $WSS=(E,C,D)$。

E 表示存储服务执行过程中存储数据节点内部计算事件，$|E|$表示服务于存储的计算数据量；

C 表示存储服务的运行业务逻辑中，与存储系统中存储设备或客户端之间的信息通信事件，其数据集可为 $\varnothing$，而 C 产生的数据量用$|C|$表示；

D 表示定义 1 中表述的宿居设备。

由于在实际应用中，用户行为、存储设备所在的环境、数据等多种因素都会影响云存储系统服务的提供，本文以主要因素为关注核心，将研究限定在数据节点可靠性与网络链路可靠性这两大不确定性因素。为了后续讨论的便利，提出以下假设。

假设 1 由于存储服务设备的物理位置相对稳定，因此设备安全性比较高，假定在工作过程中不存在任何外界自然因素对其工作性能造成影响，包括电源供应等。

假设 2 存储服务的执行质量主要依赖于存储服务设备的 CPU 与 I/O 能力。

假设 3 存储服务设备内部 I/O 能力远大于其对外网络 I/O 能力，因此服务质量瓶颈主要在于网络 I/O 能力上，且其下行带宽和上行带宽相互独立。

假设 4 存储服务设备性能和网络性能出现瓶颈而导致存储服务失效过程服从齐次泊松分布，并与其他不确定性因素出现的概率相互独立。

假设 5 存储服务对数据节点上的各硬件资源上的需求工作强度服从对数正态分布。

假设 6 节点设备上的存储服务个数服从齐次泊松分布。

假设 7 各网络连接无差异，且不考虑除网络可靠属性之外的其他不确定性因素，包括在传输过

程中的其他设备带来的影响，不考虑信号衰减与速率调整。其中，网络传输速度用 v_n 表示。

3　副本服务可靠性建模

3.1　数据节点可靠性模型

在云存储系统中，数据节点对应定义 1 所述的存储设备，根据假设 2，本研究对于数据节点性能主要关注数据节点的 CPU 与网络 I/O 能力，因此，针对这 2 个性能指标分别建立可靠性模型，最后再通过这 2 个可靠性模型构建关于数据副本服务可靠性模型。

在平均存储计算数据量为$|E|$和访问执行速度为 v_d的情况下，数据节点执行一个存储服务的平均时间 t_a 为

$$t_a = \frac{|E|}{v_d} \tag{1}$$

3.1.1　C_{cpu} 可靠模型

设 z 为单位时间存储服务期望到达数，y^k_D 指在时间 t_{a} 内节点设备 D 上同时宿居 k 个(k=0,1,⋯,K)存储服务的概率，根据假设 6，在不失一般性情况下其随机分布情况可表示为

$$y^k_D = \frac{(zt_a)^k}{k!}\mathrm{e}^{-zt_a} = \frac{\left(\dfrac{z|E|}{v_d}\right)^k}{k!}\mathrm{e}^{-\frac{z|E|}{v_d}} \tag{2}$$

设 S^k_{cpu} 表示节点设备 D 在第 k(k=0,1,⋯,K)个独立 WSS 宿居下的 CPU 失效强度，根据假设 4、假设 6，设备 D 上有 K 个服务在执行时同一时间内存储服务序列 WSS_k(k=0,1,⋯,K)为独立同指数分布的随机变量。显然，$S^k_{\text{cpu}}=f_{\text{cpu}}(cpu,k,t_a)$，其中，$f_{\text{cpu}}$ 表示与其个体 D（主要是 CPU 计算力）及存储服务 k、时间 t_a 满足一定经验函数关系。根据以上分析，C_{cpu} 不发生失效的概率可以由式(3)表示。

$$p_{\text{cpu}}(K) = \sum_{k=1}^{K}\mathrm{e}^{-S^k_{\text{cpu}}t_a}y^k_D = \sum_{k=1}^{K}\frac{\left(\dfrac{z|E|}{v_{\text{d}}}\right)^k}{k!}\mathrm{e}^{-\frac{|E|}{v_d}(S^k_{\text{cpu}}+z)} \tag{3}$$

3.1.2　N_{port} 可靠模型

N_{port} 的可靠模型建立过程与 C_{cpu} 类似。设 S^k_{port} 表示节点设备 D 在第 k(k=0,1,⋯,K)个独立 WSS 宿居下的对外网络 I/O 接口失效强度，设备 D 上有 K 个服务在执行时同一时间内存储服务序列 WSS_k(k=0,1,⋯,K)为独立同指数分布的随机变量。显然，$S^k_{\text{port}}=f_{\text{port}}(net,k,t_a)$，其中，$f_{\text{port}}$ 表示与其个体 D（主要是网络 I/O 能力）及存储服务 k、时间 t_a 满足一定经验函数关系。根据以上分析，N_{port} 不发生失效的概率可以由式(4)表示。

$$p_{\text{port}}(K) = \sum_{k=1}^{K}\mathrm{e}^{-S^k_{\text{port}}t_a}y^k_D = \sum_{k=1}^{K}\frac{\left(\dfrac{z|E|}{v_{\text{d}}}\right)^k}{k!}\mathrm{e}^{-\frac{|E|}{v_d}(S^k_{\text{port}}+z)} \tag{4}$$

3.1.3　数据节点可靠性模型

综合以上 C_{cpu} 与 N_{port} 的可靠模型，根据假设 4，C_{cpu} 不发生失效的事件与 N_{port} 不发生失效的事件相互独立，因此 $p_{\text{cpu}}(K)$与 $p_{\text{port}}(K)$对数据节点不发生失效的概率 $p_{\text{D}}(K)$符合逻辑相乘的关系，因此可以得到数据节点可靠性模型

$$p_{\text{D}}(K) = p_{\text{cpu}}(K)p_{\text{port}}(K) = \sum_{k=1}^{K}\sum_{j=1}^{K}\frac{\left(\dfrac{z|E|}{v_d}\right)^{k+j}}{k!j!}\mathrm{e}^{-\frac{|E|}{v_d}(S^k_{\text{cpu}}+S^j_{\text{port}}+2z)} \tag{5}$$

3.2　网络链路可靠性模型

根据定义 2 与假设 7，当数据节点对外提供存储服务时，数据量 C 产生的在网络中传输数据所花费的平均时间为 t_c。

$$t_c = \frac{|C|}{v_n} \tag{6}$$

令 S_{link} 表示存储服务设备与对外通信链路的失效强度，那么在通信数据集$|C|$上，该通信链路不发生失效的概率，即网络链路可靠性 p_{L} 如式(7)所示。

$$p_{\text{L}} = \mathrm{e}^{-S_{\text{link}}t_c} = \mathrm{e}^{-S_{\text{link}}\frac{|C|}{v_n}} \tag{7}$$

3.3　副本服务可靠性模型

基于以上定义与假设，结合数据节点可靠性模型与网络链路可靠性模型，显然，$p_{\text{D}}(K)$与 p_{L} 对于数据节点的副本能够正常提供存取服务的概率 $p_{\text{R}}(K)$符合逻辑相乘的关系，因此副本可靠性模型如式(8)所示。

$$p_{\text{R}}(K) = p_{\text{D}}(K)p_{\text{L}} = \sum_{k=1}^{K}\sum_{j=1}^{K}\frac{\left(\dfrac{z|E|}{v_d}\right)^{k+j}}{k!j!}\mathrm{e}^{-\frac{|E|}{v_d}(S^k_{\text{cpu}}+S^j_{\text{port}}+2z)-S_{\text{link}}\frac{|C|}{v_n}} \tag{8}$$

4 数据副本可靠分布

4.1 数据服务可靠性与副本生成

数据服务可靠性是指某一数据在整个存储系统中提供的服务可靠性，由各个副本服务的可靠性综合计算而得到，具体计算方式如式(9)所示。其中，X 为当前系统中数据副本的数量，K_x 表示当前第 x 个副本所在存储节点的访问量。式(9)显示副本数量与节点访问强度是其主要影响因素。

$$p_{\mathrm{Da}}=1-\prod_{x=1}^{X}(1-p_{\mathrm{R}}(K_x)) \tag{9}$$

为了方便判断，不失一般性地将考察的最大副本数量定为不超过 3，设相对可靠性函数 $f(K_1,K_2,\cdots,K_x)$ 为 p_{Da} 与可靠性设定阈值 $p_{\min}$ 的比，则式(10)~式(12)成立。

$$x=1,\ f(K_1)=\frac{p_{\mathrm{R}}(K_1)}{p_{\min}} \tag{10}$$

$$x=2,\ f(K_1,K_2)=\frac{1-(1-p_{\mathrm{R}}(K_1))(1-p_{\mathrm{R}}(K_2))}{p_{\min}} \tag{11}$$

$$x=3,\ f(K_1,K_2,K_3)=\frac{1-(1-p_{\mathrm{R}}(K_1))(1-p_{\mathrm{R}}(K_2))(1-p_{\mathrm{R}}(K_3))}{p_{\min}} \tag{12}$$

当系统中已存在某数据时，则其副本数量仅可能为 1、2 或 3，当系统检测到 f 函数值小于1时，说明系统中检测的数据可靠性低于阈值 $p_{\min}$，应触发新副本生成，或直接减少对该副本的访问量，或提高通信可靠性(本文暂不考虑）。当系统中不存在某数据，且系统接收到该数据初始化存储请求时，显然系统无条件触发一个新副本生成事件。这些工作将在下述副本可靠性保障机制下实现。

4.2 副本存储节点选择

在系统已经存在某数据（副本）时，根据上述检测，条件满足时将触发该数据的新副本生成，但未解决此类新副本存储于何处的问题。本文引入节点服务区域相似度的概念来辅助该问题解决。

定义 3 节点服务区域相似度

云存储系统中数据节点对外提供数据存储服务时，从满足一定存储性能上来讲，所服务的应用对象具有一定区域范围。该数据节点（原节点 u）与替换它存储服务功能的数据节点（替换节点 v）之间，应有较好的可服务对象的区域重叠性，否则这种替换将使原服务对象的请求很可能变为失效，这种节点间服务区域的重叠性称之为节点服务区域相似度（简称服务相似度），记作 $Sim(u, v)$。

设数据节点 u 服务的应用对象在一个区域 A_u 中，该区域内节点元素集合为 S_u；数据节点 v 的服务区域为 A_v，区域内节点元素集合为 S_v，并且有 $S_u\cup S_v=\{a_1,a_2,\cdots,a_M\}$，$M$ 为并集的节点总数。此外，u、v 到 $\{a_1,a_2,\cdots,a_M\}$ 中各节点的通信传输能力分别为 $\{c_{u,1},c_{u,2},\cdots,c_{u,M}\}$ 与 $\{c_{v,1},c_{v,2},\cdots,c_{v,M}\}$。则 u 与 v 间的节点服务区域相似度由式(13)计算。

$$Sim(u,v)=\frac{1}{\sqrt{\sum_{m=1}^{M}(c_{u,m}-c_{v,m})^2}} \tag{13}$$

基于服务相似度的存储节点选择方法如下。

设数据 Da 生成新副本为 R_{new}，单个副本存储点服务的可靠性阈值为 $p_{\mathrm{R},0}$，系统触发 R_{new} 生成的依据条件为 $Cond_x$，即 $f(K_1,K_2,\cdots,K_x)<1,x=1,2,3$。首先确定导致存储服务可靠性低下的关键节点，然后依次借助服务相似度和节点数据服务可靠性来选定副本存储节点。具体算法如算法 1 所示。

算法 1 副本存储节点选择算法

输入 新副本 R_{new}、新副本生成触发条件 $Cond_x$、数据节点访问服务集合 $\{K_1,K_2,\cdots,K_N\}$ 和副本存储点的可靠性阈值 $p_{\mathrm{R},0}$

输出 副本服务可靠性最低的节点 F_α、新的副本存储节点 N_{seletion}

ReplicaNodeSelection_Algorithm(R_{new}, $Cond_x$, $\{K_1,K_2,\cdots,K_N\}$,$p_{\mathrm{R},0}$)

Begin

1) 依据 $Cond_x$ 确定系统当前副本数量及相对可靠性计算公式 f;

2) 对 f 中的关联数据副本节点 α 依据式(8)计算其副本服务的可靠性值 $p_{\mathrm{R},\alpha}$;

3) 选择最低 $p_{\mathrm{R},\alpha}$ 所对应的数据副本节点作为考查节点,设为 F_α;

4) 将 F_α 所在的内层物理网络区域作为考查域 Domain(初始化 Domain 为空);

5) Repeat //确定候选存储节点 N_{seletion};

6) 在 Domain 新增域部分的节点中，选取满足节点 M_{com} 的空闲存储容量≥|Da|的所有节点构成候选节点集 S_α(不含 F_α);

7) 根据式(13)计算 F_α 与 S_α 中各元素 β 的服务相似度 $Sim(F_\alpha,\beta)$;

8) 按 $Sim(F_\alpha, \beta)$降序排列 S_α 中元素，从前部取元素(可设定) 作为筛选后的候选节点集 S_β；

9) 将 S_β 中对外网络服务能力(依剩余带宽) 低于预设值的节点排除，剩余节点构成集 S_δ；

10) 根据式(8)计算 S_δ 中各元素节点 δ 服务(假定副本存储于此) 的可靠性值 $p_{R,\delta}$；

11) 从符合 $p_{R,\delta} \geqslant p_{R,0}$ 的所有节点中选取 $p_{R,\delta}$ 最大的节点，将其作为 R_{new} 的存储节点 $N_{seletion}$；

12) 基于 F_α 所在的物理网络区域将考查域 Domain 逐步扩大；

13) Until $N_{seletion}$ 存在，或循环层数大于 3(提示“选择失败”) //确定候选节点 $N_{seletion}$ 结束

End

针对可靠性低下的节点副本，该节点选择算法从服务相似度和可靠性 2 个维度来确定一个新副本的“归属”。当然，在某数据的副本总数已经达到默认值 3 的情况下，新副本的生成（对应有新节点选择）意味着原来 3 个副本中某副本生命周期的结束，即其中可靠性低的节点副本将被删除。这将在副本分布调整算法中予以解决。

4.3 副本可靠统一分布

上述副本存储节点选择仅解决了已存在数据（副本）情况下生成新副本的存储“归属”选择问题。对于一个新来数据如何被系统初次接收存储，完成面向可靠性保障的后续多副本调整，本文提出了副本初始化存储分布与分布调整的实现算法，分别见算法 2、算法 3。其中，设 Num_R 表示数据 Da 的当前副本数量，对于数据初始化存储时该值为 0。

算法 2　副本初始化分布算法

输入　存储数据 Da、节点访问服务集合 $\{K_1,K_2,\cdots,K_N\}$、副本存储点的可靠性阈值 $p_{R,0}$

输出　候选初始化存储节点列表 $L_{seletion}$、初始化后各副本存储节点 $N_{seletion}$

ReplicaInitialization_Algorithm (D, $\{K_1,K_2,\cdots,K_N\}$, $p_{R,0}$)

Begin

1) 设置 $L_{seletion}$=空，Num_R=0，N=0；

2) 将发起数据 Da 存储请求的内层物理网络区域作为考查域 Domain；

3) Repeat　　//确定候选存储节点列表 $L_{seletion}$

4) 考查 Domain 内各数据存储节点，选取满足 M_{com} 空闲存储容量$\geqslant Da$ 的各节点构成候选节点集 S_θ；

5) 排除 S_θ 中对外网络服务能力（依剩余带宽）低于预设值的节点，将剩余节点构成集合 S_η；

6) 根据式（8）计算 S_η 中各元素节点 η 服务（假定副本存储于此）的可靠性值 $p_{R,\eta}$；

7) 选取符合 $p_{R,\eta} \geqslant p_{R,0}$ 的所有节点 η 并存入 $L_{seletion}$，对 $L_{seletion}$ 内节点按 $p_{R,\eta}$ 降序排列；

8) N=N+1；

9) 将发起 Da 存储请求的物理网络区域逐步扩大考查范围，新增区域作为 Domain；

10) Until　节点列表$|L_{seletion}| \geqslant 3$ 或 $N \leqslant 5$　//确定候选存储节点列表 $L_{seletion}$ 结束；

11) While （$Num_R \leqslant 2$ 且 $L_{seletion} \neq$空）do　//可靠性保障下的初始化存储；

12) 从 $L_{seletion}$ 中选择首元素 $N_{seletion}$，将其作为存储节点存储 Da，并从 $L_{seletion}$ 中除掉 $N_{seletion}$；

13) Num_R=Num_R+1;

14) 将 $N_{seletion}$ 上副本访问情况对应到 K_{Num_R}，并按式（10）、式（11）或式（12）计算 f (K_1, K_2,…, K_{Num_R})；

15) If　$f(K_1, K_2, \cdots, K_{Num_R}) \geqslant 1$ Then　Break End If；

16) End While//可靠性保障下的初始化存储结束

End

算法 3　副本分布调整算法

输入　新副本 R_{new}、新副本生成触发条件 $Cond_x$、数据节点访问服务集合$\{K_1,K_2,\cdots,K_N\}$、副本存储点的可靠性阈值 $p_{R,0}$

输出　已删除原副本存储节点 F_α、调整后的副本存储节点 $N_{seletion}$

ReplicaAdjustment_Algorithm(R_{new}, $Cond_x$ $\{K_1, K_2, \cdots, K_N\}$, $p_{R,0}$)

Begin

1) 依据 $Cond_x$ 确定系统当前副本值 Num_R //此时 KN 即为 K_{Num_R}；

2) 设置 M=0　//第一级调整的调整次数控制变量；

3) While $f(K_1,K_2,\cdots,K_{Num_R})<1$ 且 $M \leqslant Num_R$　Do //第一级调整（副本数不变）；

4) 执行 ReplicaNodeSelection_Algorithm，获取对应 $Cond_x$ 下副本服务可靠性最低的节点 F_α、新的副本存储节点 $N_{seletion}$；

5) 将新副本 R_{new} 存储于 N_{seletion}，并删除节点 F_α 上的副本，将 F_α 排除在副本存储节点序列外；

6) M=M+1；

7) 针对新的副本存储分布序列，并按式（10）、式（11）或式（12）计算 $f(K_1,K_2,\cdots,K_{Num_R})$；

8) End While//第一级调整结束；

9) While $f(K_1,K_2,\cdots,K_{Num_R})$< 1 且 Num_R< 2 Do //第二级调整（副本数增加）；

10) 将发起数据 R_{new} 访问请求的物理网络区域作为考查域 Domain；

11) 按照算法 2 中步骤 4)~步骤 7)完成候选节点列表 L_{seletion} 确定；

12) If L_{seletion} 为空 Then 扩大基于 R_{new} 的网络区域范围，Domain=新增域，转至步骤 11)；

End If；

13) 从列表 L_{seletion}（从前至后）选择一个节点，且该节点$\notin$现有副本存储节点序列，将其作为新增的副本存储节点 N_{seletion}；

14) 将新副本 R_{new} 存储于 N_{seletion} 中，并从 L_{seletion} 中除掉 N_{seletion}；

15) Num_R=Num_R+1；

16) 针对新的副本存储分布序列，并按式（10）、式（11）或式（12）计算 $f(K_1,K_2,\cdots,K_{Num_R})$；

17) End While//第二级调整结束

18) If 当前 $f(K_1,K_2,\cdots,K_N)$<1 且 $Num_R\geqslant$ 3 Then

19) 输出“副本默认值为 3 时，可靠性保障下不能完成调整”，退出本算法；

20) End If

End

副本统一分布是为可变数量的副本统一确定节点存储，核心是根据副本可靠性、数据访问量和网络通信等情况来完成副本的初始分布和分布调整。分布调整关联的前驱过程有：数据服务可靠性低于阈值 $p_{\min}$ 判别、新副本生成、基于服务相似度的存储节点选择等。分布调整分 2 个级别进行：“副本数不变的调整”和“副本数增加的调整”；在保持系统默认副本数量前提下优先实现第一级调整。

依据副本分布算法，可能带来大量的新副本生成并且其存储于新数据节点的情况。随着这类存储服务过程的持续，系统可能产生不必要的副本冗余，这将涉及可靠性保障下的副本删除。假设数据 Da 在节点 u 所支持的访问量为 K_u^D，则若将该副本删除，该副本的访问量将被分配到同源数据的其他副本所在节点（设该类节点为 v，其数目为 V）上。因此，为了减少对原有数据存储总服务的影响，这种访问量的合理分配转移，可以依据节点 u 与其他各节点 v 之间的服务相似度，即选择具有较大 $Sim(u,v)$的节点 v 来分担节点 u 更多的访问任务。设 $\Delta K_{u\to v}^D$ 为 u 删除后适宜转移给 v 的访问服务量比率，可按照式(14)求解，并有式(15)成立。

$$\Delta K_{u\to v}^D=\frac{Sim(u,v)}{\sum_{v=1}^{V}Sim(u,v)}K_u^D,\quad v\neq u \tag{14}$$

$$K_V^D=K_V^D+\Delta K_{u\to V}^D,\quad v\neq u \tag{15}$$

为了有效考查存储节点的访问量变化，系统设定某数据副本访问量考查周期为 T（如 1 天），再依据数据的先验周期访问量、数据副本服务可靠性等条件进行节点中副本的可删除性判断与删除。本文设计副本删除算法，如算法 4 所示。

算法 4 副本删除算法

输入 系统中所有存储数据的数据项列表 L_{data}、各存储数据 Da 所在的数据节点的访问量集合 $\{K_1,K_2,\cdots,K_N\}$、副本存储点的可靠性阈值 $p_{\text{R},0}$

输出 已删除副本的原存储节点 u、删除后的副本存储节点列表 L_{DNode}

ReplicaDeletion_Algorithm (L_{data}, $\{K_1,K_2,\cdots,K_N\}$,$p_{\text{F},0}$)

Begin

1) 对 L_{data} 列表中的各数据项 Da 建立它的副本存储节点列表 L_{DNode}（副本$|L_{\text{DNode}}|\leqslant$3）；

2) While 从 L_{data} 中成功选取一个未曾选取过的数据项 Da Do //对 L_{data} 中各数据项进行判断；

3) 设 Num_R=$|L_{\text{DNode}}|$，K_{mean} 为数据 Da 的先验周期平均访问量；

4) If 连续 5 个 T 内 Da 的每个周期访问量皆小于 $K_{\text{mean}}/2$ 且 $Num_R\neq$1 且$(K_1,K_2,\cdots,K_{Num_R})$>1 Then //实施节点删除的条件判断；

5) While 从 L_{DNode} 中成功选取一个未曾选取过的副本存储节点 u Do //穷举列表中节点以试探；

6) 拟删除 u 上副本，对 L_{DNode} 中其余各节点 v（$u\neq v$），按式（14）、式（15）求解 $\Delta K_{u\to V}^D$ 和 K_v^D；

7) Da 形成新的数据节点的访问量集合 $\{K_1,K_2,\cdots,K_{Num_R-1}\}$ //所有 v 节点的数目=(Num_R−1)；

8) 按照式（10）、式（11）或式（12）计算相

对可靠性函数 f 的值；

9) If $f(K_1,K_2,\cdots,K_{Num_R-1})\geqslant 1$ Then　//$f\geqslant 1$ 时实施删除，并继续试探其他删除；

10) 删除 u 上数据 D 的副本；

11) 将 u 从 L_{DNode} 中移除；

12) $Num_R=Num_R-1$；

13) End If　//隐含 $f<1$ 时不满足可靠性而不删除 u，并继续试探其他删除；

14) End While　//穷举列表中节点以试探结束

15) End If　//实施节点删除的条件判断结束

16) End While　//对 L_{data} 中各数据项进行判断结束

End

通过上述 4 个算法，系统结合可靠性模型将根据存储数据节点的访问量等因素计算该副本及数据服务的可靠性值，并最终确定新副本的生成时机，完成副本的存储节点选择、统一分布和副本删除。本方法避免了系统受限于固定的数据副本数量，而是根据实际情况来增加与删除副本。

5　系统实现与相关实验

ERS-Cloud 是研究团队利用开源平台 Hadoop 的 HDFS 搭建起来的教育资源云存储系统，该系统通过增设属性 strategyStatus 而动态变更存储策略，若该值为 0，则采取系统默认的存储管理（含副本放置）机制；若为 1，则采用本文提出的基于可靠性模型的副本分布策略。平台依托操作系统 Ubuntu 10.04，基于 Hadoop 0.21.0 源码进行改进，Web 服务器为 Tomcat 6.0.30，云平台主要是由 1 台 Web 服务器、1 台 NameNode 主机与 11 台 DataNode 主机构成。系统中存储高校课程等各类资源 1.23 TB。实验邀请了华南师范大学某学院全体师生参与为期一月的资源数据应用测试，模拟云存储的真实应用环境。根据同类研究多采用与默认副本机制进行对比的实验方法，本文设计了如下 3 组实验：与默认机制对比下的存储服务性能比较；与默认机制、PRCR 机制 [10]对比下的数据空间利用状况比较；本研究中的副本误删率测试。实验中将 Hadoop 默认的副本存储方法简称策略 1，本文所提出的副本可靠性保障方法简称策略 2，实验分述如下。

5.1　系统服务性能测试

该部分实验主要对比 2 种策略下系统中数据的可靠性程度，及其在不同访问模式下系统的资源数据访问效率，具体实验结果如图 1 和图 2 所示。

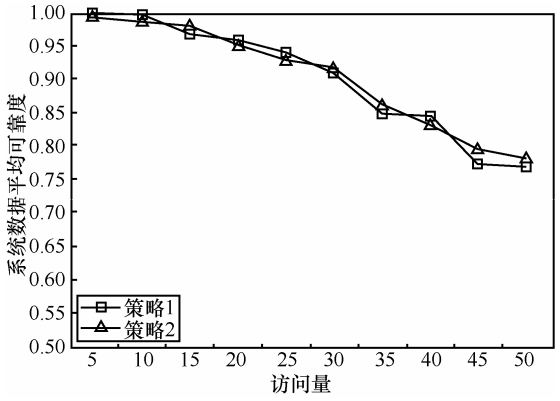


图 1　2 种策略下系统存储数据的平均可靠性对比

从图 1 可以看出，2 种策略下系统中数据的可靠性程度并没有太大的差异性，这说明了策略 2 能够在相比策略 1 之下减少副本数量的同时，确保数据的可靠性，但随着单位时间用户访问量的提高，2 种策略的管理下的数据可靠性均下降。

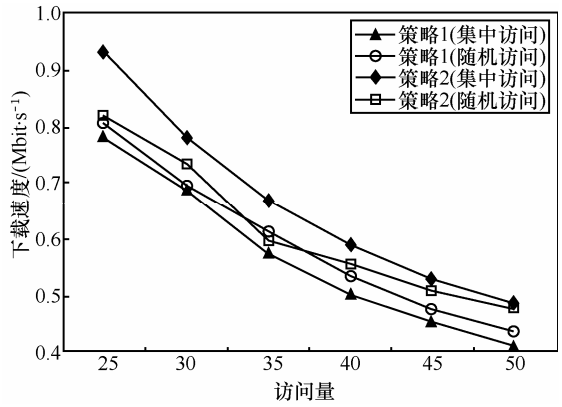


图 2　2 种策略下资源下载速度对比

图 2 说明了随着访问量的增大，4 种情况的下载速度都明显下降；当用户访问相对集中时使用策略 2 的传输效率最高；资源被集中访问现象不明显时 2 种策略的下载速度差异不大，但随着用户量增大采用策略 2 的下载速度较策略 1 更快；策略 1 下资源被集中访问现象明显时下载速度最低。由此看来，当用户的访问相对集中于局部数据时，策略 2 能够有效提高资源的访问效率。

5.2　数据空间利用状况测试

资源数据空间利用状况测试首先是通过调节用户访问量、策略 2 中数据可靠性最低阈值 $p_{\min}$ 来检测系统中资源数据总体大小的变化，并与 Hadoop 的静

态副本存储机制进行对比，实验结果如图 3 所示。鉴于文献[10]中提出的 PRCR 机制也以控制副本数量为关注点，这与本研究具有某些相似性，因此，为了进一步测试相关指标，本实验也将与 PRCR 机制进行比较，对比指标为系统中不同副本数量（副本数量为 1、2 与 3）的文件所占系统文件总数的比例（副本数与文件数的比值）与平均副本数，其中，PRCR 的比例值则是参考文献[10]中的实验数据 1.6:9（即系统中副本数量为 2 与副本数量为 1 的数据量比值）。实验结果如图 4 所示。

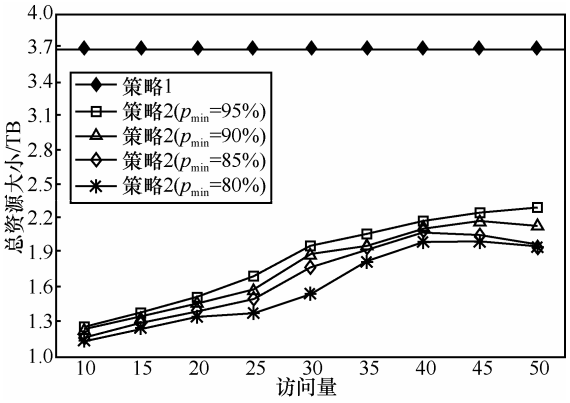


图 3　2 种策略下系统的数据存储利用空间对比

由图 3 知策略 2 能够相比策略 1 大大减小系统冗余存储所占的空间，但随着系统单位时间访问量的不断提高，为了确保访问数据的可靠性，一些访问率较高的节点上的数据逐渐增加副本数量，造成了总资源大小的提高。显然，当最低可靠性阈值 $p_{\min}$ 越小时，副本的增加速度越缓慢。由于本实验中 Hadoop 采用的是静态副本技术，默认副本数量是 3，为此资源存储总量大小没有发生变化。

从图 4 可知，PRCR 的 1+2 存储机制中副本的数量最多为 2，而当策略 2 的 $p_{\min}$ 值为 95%时，平均副本数量比 PRCR 高，而当 $p_{\min}$ 值为 90%、85%、80%时，策略 2 的平均副本数量越来越小，且比 PRCR 低。而由于 2 种方法中对数据可靠性的计算方式不同，两者的可靠性值不宜直接进行对比，策略 2 的可靠性计算值相对较低，95%是对云存储系统中数据可靠性要求较高，因此图 4 可以说明策略 2 相比 PRCR 的 1+2 存储机制在数据副本数量的控制具有一定程度的改善。

5.3　副本误删率测试

由于副本删除算法是基于对副本访问量、节点存储情况等的预测，而它们都有瞬时发生较大变化的可能性，因此存在一定的预测不准确性，这将导致数据副本的误删。因此，实验拟通过调节数据（副本集合）最低可靠性阈值 $p_{\min}$ 来检测副本删除算法是否会对系统造成不良影响。结果如图 5 所示。

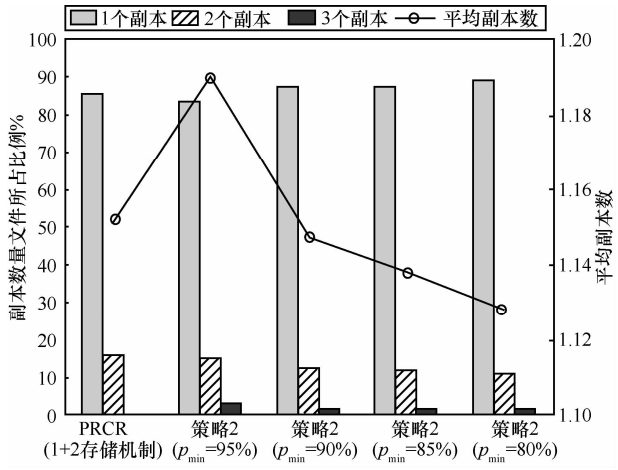


图 4　不同副本数量比例图与平均副本数关系

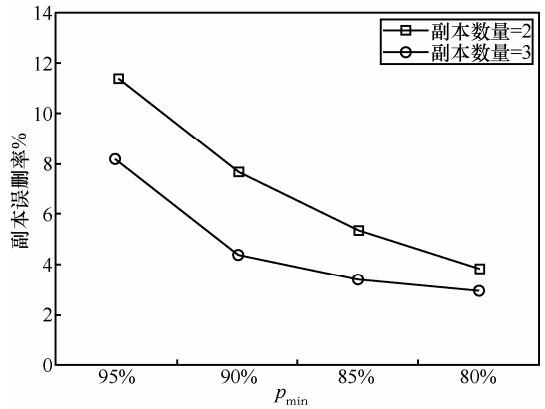


图 5　副本数量与副本误删率关系

从图 5 可以看出，随着最低可靠性阈值 $p_{\min}$ 变小，副本的误删率也随之变小，且当副本数量为 3 时，误删率较副本数量为 2 时更小，但误删率均在可以接受的范围内。2 种副本数量情况的误删率较小说明了副本删除算法具有一定的稳定性。

综合以上实验结果，可以得知本文提出的副本分布方法能够保证数据服务的可靠性，且效果与系统默认 3 个存储副本时的情况基本相当，但能够较为明显地减少副本冗余造成的存储资源浪费，在资源集中访问的情况下存储访问服务效率良好；基于副本删除算法的副本误删率较低，相应的恢复机制担当任务较轻，减少了不必要的工作消耗。

6　结束语

为了进一步提高云存储系统的资源利用效能，分析存储节点本身及其访问情况，构建副本、数据服务可靠性模型，该模型能够表征副本服务可靠性与副本数量、存储节点的访问量等之间的关系。提出了副本生成时机的判断方法，实现副本存储节点选择算法、副本可靠分布算法、副本删除算法，通过应用系统进行实验，效果显示，基于可靠性模型的副本管理策略能对云存储系统应用效果保障与冗余度降低发挥重要作用。如何使资源云存储应用系统提供稳定的服务质量是一个内涵丰富的主题，它涉及多方复杂因素，解决方法多样，且多与应用服务领域紧密相关，本团队下阶段拟基于应用特征建模研究动态存储支持策略。针对本研究中的云存储副本生成、分布、删除算法，计划立足环境感知进行优化。云存储中数据的正确性保持、多副本的一致性检测与维持也是未来研究的重点工作。

参考文献：

[1] GHEMAWAT S, GOBIOFF H, LEUNG S T. The google file system[J]. ACM SIGOPS Operating Systems Review, 2003,37(5): 29-43.

[2] SHVACHKO K, KUANG H, RADIA S, *et al*. The hadoop distributed file system[A]. Proc of IEEE MSST 2010[C]. Incline Village, NV, USA, 2010. 1-10.

[3] WEI Q S, VEERAVALLI B, GONG B Z, *et al*. CDRM: A cost-effective dynamic replication management scheme for cloud storage cluster[A]. Proc of IEEE CLUSTER 2010[C]. Heraklion, Greece, 2010. 188-196.

[4] BONVIN N, PAPAIOANNOU T G, ABERER K. A self-organized, fault-tolerant and scalable replication scheme for cloud storage[A]. Proc of ACM SOCC 2010[C]. Indianapolis, IN, USA, 2010.205-216.

[5] LIAO B, YU J, SUN H, *et al*. A QoS-aware dynamic data replica deletion strategy for distributed storage systems under cloud computing environments[A]. Proc of IEEE CGC 2012[C]. Xiangtan, China, 2012. 219-225.

[6] 宋娅菲. 基于竞标模式的云存储副本放置策略研究[D]. 武汉,中国，华中师范大学, 2012.
SONG Y F. Research on Replica Placement Strategy in Cloud Storage Based on Bidding Model[D]. Wuhan, China, Central China Normal University, 2012.

[7] HE R, LUAN Z Z, HUANG Y Q, *et al*. Providing high availability for distributed stream processing application with replica placement[A]. Proc of IEEE NBiS 2012[C]. Melbourne, Australia, 2012.685-690.

[8] WU S C, CHEN G Z, GAO T G, *et al*. Replica pre-adjustment strategy based on trend analysis of file popularity within cloud environment[A]. Proc of IEEE CIT 2012[C]. Chengdu, China, 2012.219-223.

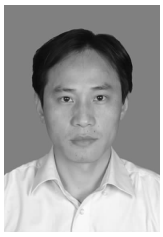
[9] XU X D, WANG S X, YAO K B, *et al*. Research on the strategy of FLDC replication dynamically created in cloud storage[A]. Proc of IEEE CECNet 2012[C]. Yichang, China, 2012.2815-2818.

[10] LI W H, YANG Y, CHEN J J, *et al*. A cost-effective mechanism for cloud data reliability management based on proactive replica checking[A]. Proc of IEEE/ACM CCGrid 2012[C]. Ottawa, ON, USA, 2012.564-571.

[11] FENG Q Q, HAN J Z, GAO Y, *et al*. Magicube: high reliability and low redundancy storage architecture for cloud computing[A]. Proc of IEEE NAS 2012[C]. Xiamen, China, 2012.89-93.

[12] BAUER E, ADAMS R. Reliability and Availability of Cloud Computing[M]. Hoboken, NJ, USA: Wiley-IEEE Press, 2012.

作者简介：



黄昌勤（1972-），男，湖南常德人，博士，华南师范大学教授、博士生导师，主要研究方向为可信云服务、语义智能及其教育应用。

李源（1987-），男，广东汕头人，华南师范大学硕士生，主要研究方向为云存储、教育资源管理。

吴洪艳（1970-），女，黑龙江齐齐哈尔人，华南师范大学博士后，主要研究方向为适应性学习系统。

汤庸（1964-），男，湖南张家界人，博士，中山大学教授、博士生导师，主要研究方向为大数据与时态信息处理、协同计算。

罗旋（1990-），男，湖南桃源人，华南师范大学硕士生，主要研究方向为云计算、社会性软件。