第34卷 第17期 计 算 机 工 程 2008年9月
Vol.34 No.17 Computer Engineering September 2008
·网络与通信· 文章编号：1000—3428(2008)17—0088—03 文献标识码：A 中图分类号：TP393

高可靠性元数据服务器研究

刘 群[1]，冯 丹[2]，王 芳[2]
(1. 华中科技大学网络与计算中心，武汉 430074；2. 华中科技大学信息存储系统教育部重点实验室，武汉 430074)

摘 要：基于对象存储的核心是将存储空间管理从存储应用中分离，其中元数据服务器(MDS)负责逻辑视图管理和提供全局命名空间。该文针对MDS的引路导航性能，提出一种主从备三重链式结构，在不增加硬件成本的前提下，能提供无间断的MDS服务，确保了系统的高可靠性。

关键词：元数据服务器；可扩展对象的海量存储系统；高可靠性；主从备链式结构

Research on MetaData Server of High Reliability

LIU Qun[1], FENG Dan[2], WANG Fang[2]
(1. Network and Computing Center, Huazhong University of Science and Technology, Wuhan 430074;
2. Key Laboratory of Data Storage System, Ministry of Education, Huazhong University of Science and Technology, Wuhan 430074)

【Abstract】 The most immediate effect of object-based storage is the offloading of space management from storage applications. MetaData Server (MDS) is responsible for the logical view management and provides unique naming space. To play an important role of navigation for MDS, this paper proposes a method of master/slave/standby chain backup. It can provide a continuous serving and ensure high reliability of MDS without adding any hardware device.

【Key words】 MetaData Server(MDS); Based on Scalable Object Mass Storage System(BSO-MSS); high reliability; master/slave/standby chain

1 概述

随着网络技术与存储技术的发展结合，出现了存储区域网(Storage Area Network, SAN)和附网存储(Network Attached Storage, NAS)两种系统。SAN将元数据与数据进行分离，采用专用的元数据服务器(MetaData Server, MDS)管理系统中的所有元数据，用户通过网络可直接访问“块”设备。NAS将存储设备直接连接到网络中，供“文件”级访问接口，文件服务器只提供网络文件协议便于共享，由NAS自行管理元数据，并处于数据访问路径上。在大规模的存储系统中，尽管元数据的数据量相对于整个存储系统而言比较小，但统计表明，在文件系统的访问中，对元数据的访问次数占全部访问的50%~80%[1]。因此，随着当今存储需求的日益增长，数据量会急剧增加，SAN和NAS的元数据访问都存在瓶颈。

一种新兴的对象存储(Object Based Storage, OBS)[2]体系结构成为存储研究领域的热点。OBS将存储数据的逻辑视图和物理视图分开，MDS仅管理约占负载10%的文件管理部分，而剩下约90%的存储管理部分下移到基于对象的存储设备(Object-based Storage Device, OSD)中，OSD负责其中所有对象的空间分配(对象映射的逻辑块)，维护与对象有关的元数据(类似于Unix/Linux的inode结构)，对象的存储建立在一个“平坦”一维空间中，接口也从基于块的接口变为基于对象的接口，这样与存储管理相关的元数据分布到整个系统各个OSD中。因此，OBS克服了NAS和SAN中不足，既有“块”接口的快速，又有“文件”接口的便于共享，并在扩展性、安全性、高性能和跨平台数据共享等方面更胜一筹。

MDS一般分为集中式和分布式2种方式的服务。集中式MDS服务采取专用的MDS管理元数据，其具有共享程度高、实现相对简单等优点，主要采用主/从备份方式，但可靠性较差，如对象存储系统Lustre[3]、Panasas[4]和UCSC[5]等都属于这种方式，一旦2台MDS同时出现故障，系统就无法正常使用。分布式MDS服务则是系统中没有专用的MDS，元数据分散存放在存储设备中，所有存储设备既可存放元数据，又可存储用户数据，这种方式能提供高性能和高可扩展性，但系统非常复杂，并辅以协作cache、分布式锁机制、冗余、校验等技术来满足可靠性。本文以可扩展对象的海量存储系统(Based on Scalable Object Mass Storage System, BSO-MSS)为背景，采用集中式MDS服务方式，提出一种主从备三重链式结构，对MDS的可靠性服务进行了讨论与分析。

2 扩展对象的海量存储系统

扩展对象的海量存储系统是由MDS集群、存储对象(Storage Object, SO)和客户端组成。

MDS如同一盏导航灯，提供全局的命名空间，负责整个系统的目录结构、权限控制和文件命名。它建立目录与文件的访问管理，构建一个树状目录结构，包括目录和文件的创建、删除和访问控制，以及限额控制等；同时为客户端提供统一的文件逻辑视图，构造、管理并描述每个文件的分布视图，允许客户端直接访问该文件数据。

SO是一个智能设备，它包括处理器、内存、网络接口和

基金项目：国家“973”计划基金资助项目“下一代互联网信息存储的组织模式和核心技术研究”(2004CB318201)；华中科技大学校实验技术研究基金资助项目

作者简介：刘 群(1969－)，女，工程师、博士，主研方向：计算机网络，高性能网络存储；冯 丹、王 芳，教授、博士

收稿日期：2008-02-28 **E-mail**：liliuqun@mail.hust.edu.cn

存储设备，其中，存储设备以多种形式存在，可以是 Disk, RAID, NAS 和磁带设备等[6]。依照 SO 的接口与状态，实现数据组织的存储。SO 与传统的存储设备区别不是介质，而是接口，并且在 OBS 基础上进一步丰富和扩展了对象接口的内涵，使之成为一个通用存储访问的接口模式。客户端的应用程序采用 POSIX 方式进行数据访问。

在 BSO-MSS 中，MDS 为所有客户端提供了一个全局的、有效的、单一的目录树，目录树包含从不同 MDS 移植到这个命名空间来的目录和文件，并提供 POSIX 标准文件访问接口。同时针对元数据访问的特点，设置缓冲区，将最近访问过的元数据缓存到该缓冲区中，减少磁盘 I/O。由此可见，MDS 的元数据服务在 BSO-MSS 中起着举足轻重的作用。

3 高可靠的元数据服务方法

3.1 可靠性方法

客户端访问 BSO-MSS 时，都必须先访问 MDS，获得相应的对象信息。因此，随着海量存储系统规模的不断扩大，访问频率也急剧增长，加大了 MDS 失效的可能。为了保证 MDS 提供连续不间断的元数据服务，系统应具有高可靠性。

防止错误发展成为故障(即容错)是提高可靠性的主要方法，典型容错方法有镜像复制和冗余机制 2 种。镜像复制技术是把数据块快速备份到另一个设备中，通常采用 2-路复制和 3-路复制，其中，2-路复制可分为主/从模式和主-从/主-从模式。主/从模式是指系统中主 MDS 负责元数据服务，而具有相同数量的从 MDS 长期处于后备的监控状态，通过镜像及时备份数据。当“心跳”信号出现异常时，表明主 MDS 发生故障，或者从 MDS 无法收到主 MDS“心跳”信号，管理软件则指令主 MDS 停止工作，将系统资源转移到从 MDS 上，从 MDS 替代主 MDS 工作，这样在短时间内保证元数据服务完全恢复正常使用。

主-从/主-从模式是系统中所有 MDS 既是主 MDS，又是从 MDS，提供了双向故障恢复功能，当某一台 MDS 出现故障时，另一台 MDS 作为它的从 MDS 在短时间内将主 MDS 的接管过来，从而保证了元数据服务的持续性。虽然这种方式将 MDS 总数目减少一半，但若有一台 MDS 失效进行切换后，另一台 MDS 就会同时响应原来由两台 MDS 响应的请求，负载过大，有可能造成元数据服务的性能瓶颈。

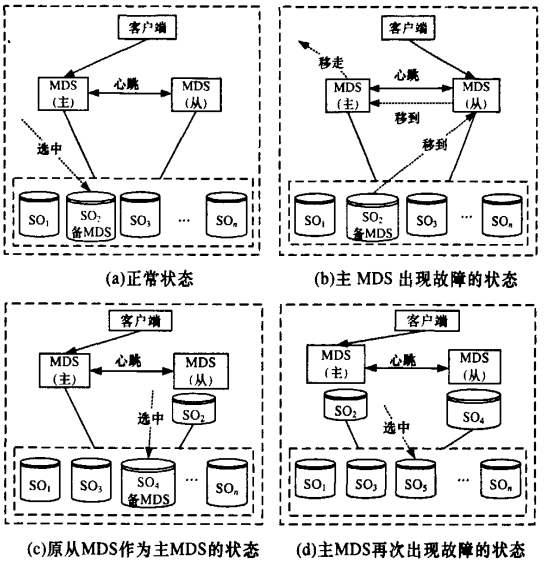
3-路复制采用主、从、备模式，有一台主 MDS 负责服务，一台 MDS 和一台 MDS 长期处于后备的监控状态，通过镜像及时备份数据。虽然与 2-路复制相比，大大提高了系统可用性和可靠性，但从系统资源方面考虑，存在很大的浪费，性价比随着系统扩展而急剧降低，一般很少采用。

冗余机制采用冗余数据实现系统纠错和提高可靠性，通过并行访问实现高性能，例如 RAID5。但当系统对多台 MDS 的共享元数据并发请求访问，易形成 I/O 瓶颈，从而降低系统性能。同时当磁盘失效时，重建新磁盘需要花费相当长的时间，这使在数据恢复期间有可能发生另外一个磁盘失效。假设单台 MDS 的可用度是 99.9%，对于 2 台 MDS 对称共享 RAID5，可靠系数则是 99.95%；而采用镜像复制的主/从模式 MDS 服务可靠系数为 $1-(1-99.9\%)^2=99.999\%$，显而易见，镜像复制技术的可靠性高于冗余机制技术。

3.2 三重链式结构

在 BSO-MSS 中，随着系统的扩展，SO 在硬件上有所不同。同时由于对象分配策略的不同，也会造成各个 SO 的负载不均衡，并由其属性反映。因此，本文在主/从备份模式基础上，提出一种主从备三重链式 MDS 服务方法，系统仍然采用两台专用 MDS——主 MDS 和从 MDS。通常情况下，由主 MDS 对外提供服务，从 MDS 通过心跳监测主 MDS 状态，同时根据当前系统中 SO 的属性，选择负载最轻的 SO 作为备用 MDS。

当主 MDS 出现故障时，从 MDS 启动并替代主 MDS 提供向外服务，此时主 MDS 可从系统中移走进行修复，备用 MDS 转变为从 MDS，与此同时，在余下的 SO 中选择负载最轻的作为备用 MDS，依次类推，形成一种三重链式结构，如图 1 所示。



(a)正常状态　(b)主 MDS 出现故障的状态

(c)原从MDS作为主MDS的状态　(d)主MDS再次出现故障的状态

图 1　主从备三重链式结构

此外，专用主 MDS 的优先级最高，其次是专用从 MDS，也就是说，当出现故障时，首选专用 MDS，特别是专用主 MDS 一旦修好，立即转化为由专用 MDS 担当主 MDS，当且仅当 2 台专用 MDS 都同时出现故障时，才由从 SO 中选取 MDS 提供真正的服务，这样既不增加系统的硬软件成本，又充分利用现有的资源。

3.3 可靠性评价

在 BSO-MSS 中，MDS 担负着导航引路和文件与对象映射等的重要作用，为了避免由于它的失效造成无法挽回的巨大损失，需对它进行有效准确的可靠性评价，成为一项必不可少的环节。

对元数据服务进行可靠性分析，首先需对元数据服务的各种可靠性性能指标进行量化。本文采用目前应用广泛并具代表性的性能指标，如失效率λ、修复率μ、可用度 A、平均无故障时间(Mean Time To Failure, *MTTF*)、平均修复时间(Mean Time To Repair, *MTTR*)[7]，以及平均服务时间(Mean Time To Service, *MTTS*) 等。其中，根据 *MTTF* 和 *MTTR* 的定义，可得 $MTTF=1/\lambda$, $MTTR=1/\mu$。

马尔可夫链(Markov Chain, MC)模型是一个广泛用于评价系统可靠性的工具。MDS 服务过程是指某一个 MDS 从正常工作状态转移到失效状态，然后经过修复回到正常服务状态，如此往复循环。因此，MDS 的状态 M 分为 2 类：$M=S\cup F$，其中，S 表示 MDS 正常服务状态集合；F 表示 MDS 失效集合。

为了便于分析，本文作如下假设：

(1)系统内所有节点的状态改变是瞬间完成的，这样考虑可以避免 MC 模型中出现太多对结果影响不大的状态。

(2)不同节点的失效概率和故障修复概率是相互独立的，不会因为某个节点的失效而导致与其相关的节点的失效概率有所增加。

主从备三重链式结构是采用镜像复制的双重冗余技术，系统状态定义如下：

(1)状态 0：主从备 MDS 均正常，MDS 能提供正常服务；

(2)状态 1：主 MDS 失效，从、备 MDS 两台正常，MDS 能提供正常服务；

(3)状态 2：主、从 2 台 MDS 失效，备 MDS 正常，MDS 能提供正常服务；

(4)状态 3：主从备 3 个 MDS 均失效，MDS 不能提供正常服务。

图 2 为主从备三重链式的 MDS 服务系统状态转移图。

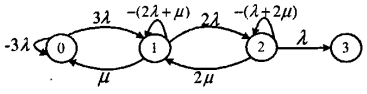


图 2 MDS 状态转移

状态 0～2 为 MDS 正常服务状态，S={0,1,2}；状态 3 为 MDS 失效状态，不能提供 MDS 服务，F={3}；λ和μ分别表示系统中 MDS 的失效率和修复率，即λ=1/*MTTS*，μ=1/*MTTR*。若只考虑可靠状态集合 S，则可以通过式(1)和式(2)计算出系统平均服务时间 *MTTS*。

$$\boldsymbol{T}_S\boldsymbol{Q}_S=-\boldsymbol{P}_S(0) \tag{1}$$

$$MTTS=\sum_{i\in S}T_i \tag{2}$$

其中，$\boldsymbol{T}_S$为某一服务状态到达失效状态前的平均时间向量，也就是该状态的平均工作服务时间向量；$\boldsymbol{Q}_S$为某一服务状态的转移概率矩阵；$\boldsymbol{P}_S$为初始状态概率向量；*MTTS* 为系统平均服务时间。因此，正常服务状态 S 的状态转移矩阵为

$$\boldsymbol{Q}_3=\begin{bmatrix}-3\lambda & 3\lambda & 0 & 0\\ \mu & -(2\lambda+\mu) & 2\lambda & 0\\ 0 & 2\mu & -(\lambda+2\mu) & \lambda\\ 0 & 0 & 0 & 0\end{bmatrix}$$

通过式(1)和式(2)计算 MDS 在失效状态前平均服务时间 *MTTS*[8]。

$$MTTS=\frac{11\lambda^2+7\lambda\mu+2\mu^2}{6\lambda^3} \tag{3}$$

则系统可用性为

$$A=1-(1-\frac{MTTS}{MTTS+MTTR})^2=1-(\frac{6\lambda^3}{6\lambda^3+11\lambda^2\mu+7\lambda\mu^2+2\mu^3})^3 \tag{4}$$

BSO-MSS 的当前参数定义见表 1。

表 1 系统相关参数及意义

参数	参数值	意义
$1/\lambda$	10^5 h	单个硬盘平均无故障时间
μ	5 GB/h	故障修复时间

假设单个 MDS 与 SO 只有一块硬盘，则其平均服务时间 *MTTS* 为 100 h，可用性 A 为 0.998，将现有的镜像复制机制与本文的提出的主从备三重链式结构在可靠性和性价比方面进行比较，结果见表 2，其中，N 为专用 MDS 数量。

表 2 可靠性与性价比的比较

参数	*MTTS*/h	*A*	*N*
单个 MDS	100	0.998	1
2 路-复制主/从 MDS	25 150	$1-1.58\times10^{-11}$	2
3 路-复制 MDS	8 391 800	$1-1.35\times10^{-23}$	3
主从备三重链式 MDS	8 391 800	$1-1.35\times10^{23}$	2

SO 在数量上远远超过 MDS，同时对象分配策略的不同，某一时刻时所有 SO 的负载均可不同，利用对象属性特性了解当前 SO 的负载，选择当时负载最轻的 SO 作为备用 MDS，形成主从备三重链式结构，这样不仅没有提高硬件成本，而且在可靠性和可用性更胜一筹。虽然需要增加一些存储开销，但能够分担系统的访问负载，也就是说，当访问 MDS 存在热点时，可实施访问迁移，实现负载均衡，同时提高了系统的吞吐率。

4 结束语

本文通过分析 NAS 和 SAN 中不足，以及 OBS 的特性，在 BSO-MSS 的背景下，针对 MDS 的导航性，以容错作为切入点，提出一种主从备三重链结构 MDS 服务方法，它不需要增加额外的设备，充分挖掘了存储对象节点的富余资源，选取负载最轻的节点来承担 MDS 功能，这样保证了系统高可靠性和可用性，而且避免了热点访问。

参考文献

[1] Ousterhout J K, Costa H D, Harrison D, et al. A Trace-driven Analysis of the Unix 4.2 BSD File System[C]//Proceedings of the 10th ACM Symposium on Operating Systems Principles. [S. l.]: ACM Press, 1985.

[2] Mesnier M, Ganger G R, Riedel E. Object-based Storage[J]. IEEE Communications Magazine, 2003, 41(8): 84-90.

[3] Braam P J. The Lustre Storage Architecture[Z]. 2004-03.

[4] Panasas Inc. Object-based Storage: Enabling Peta-scale Computing [EB/OL]. (2003-11-01). http://www.panasas.com/docs.

[5] Hospodor A, Miller E L. Interconnection Architectures for Petabyte-scale High-performance Storage Systems[C]//Proceedings of the 21st IEEE/12th NASA Goddard Conference on Mass Storage Systems and Technologies. College Park, MD, USA: [s. n.], 2004.

[6] Liu Qun, Feng Dan, Qin Ling-jun, et al. A Framework for Accessing General Object Storage[C]//Proceedings of the 2006 International Workshop on Networking, Architecture, and Storages. Dalian, China: [s. n.], 2006.

[7] 张 涛，胡东成. TCRES: 一种实用的容错计算机系统可靠性评价系统[J]. 计算机研究与发展, 1997, 34(9): 716-720.

[8] Baek S H, Kim B W, Joung E J, et al. Reliability and Performance of Hierarchical RAID with Multiple Controllers[C]//Proceedings of the 20th Annual ACM Symposium on Principles of Distributed Computing. New York, USA: ACM Press, 2001.

高可靠性元数据服务器研究

作者： 刘群， 冯丹， 王芳， LIU Qun， FENG Dan， WANG Fang

作者单位： 刘群,LIU Qun(华中科技大学网络与计算中心,武汉,430074)， 冯丹,王芳,FENG Dan,WANG Fang(华中科技大学信息存储系统教育部重点实验室,武汉,430074)

刊名： 计算机工程 ISTIC PKU

英文刊名： COMPUTER ENGINEERING

年，卷(期)： 2008，34(17)

被引用次数： 0次

参考文献(8条)

1. Ousterhout J K. Costa H D. Harrison D A Trace-driven Analysis of the Unix 4.2 BSD File System 1985

2. Mesnier M. Ganger G R. Riedel E Object-based Storage 2003(08)

3. Braam P J The Lustre Storage Architecture 2004

4. Panasas Inc Object-based Storage:Enabling Peta-scale Computing 2003

5. Hospodor A. Miller E L Interconnection Architectures for Petabytescale High-performance Storage Systems 2004

6. Liu Qun. Feng Dan. Qin Ling-jun A Framework for Accessing General Object Storage 2006

7. 张涛. 胡东成 TCRES:一种实用的容错计算机系统可靠性评价系统[期刊论文]-计算机研究与发展 1997(09)

8. Baek SH. Kim B W. Joung E J Reliability and Performance of Hierarchical RAID with Multiple Controllers 2001

相似文献(1条)

1. 学位论文 刘群 基于可扩展对象的海量存储系统研究 2006

信息存储是人类社会永恒的需求。随着计算机技术的发展和应用的普及，信息存储容量成爆炸性地增长，现有网络存储系统已无法满足人们对于存储的需要。基于对象存储（Object-Based Storage，OBS）技术逢时崛起，利用现有的存储组件、处理技术和网络技术，通过简单方式来获得前所未有的高吞吐量，成为下一代网络存储的主流。它采用包含数据和属性的“对象”作为接口，既有了“块”接口的快速，又有“文件”接口的便于共享，并分离了存储数据的逻辑视图和物理视图，将存储数据的逻辑视图保留在元数据服务器中，而物理数据存放在基于对象存储设备（Object-Based Storage Device，OSD）中。同时，它将传统文件分解为系列数据对象，分发到一个或多个OSD 中。虽然对象给存储系统带来了一种新的理念，但现有的与对象相关的存储系统中对象都仅定义为非定长的数据单位，束缚了“对象”这个有着丰富内涵的词汇。

基于可扩展对象的海量存储系统（Based on Scalable Object Mass StorageSystem，BSO-MSS）吸取了OBS 的优点，在“对象”现有的含义基础上扩充，使它不仅只包括用户数据，还将目录、文件、存储设备管理等纳入对象之中，形成层次结构的对象体系结构，实现对象的分布存储、层次管理的模式，并建立基于存储对象统一访问模式，将块、对象和文件三种存储接口进行融合与统一。这样不仅具有统一逻辑视图、数据共享、主动服务、并行访问、统一存储和易管理等特点，而且有着其他存储结构难以达到的高可扩展性和高性能。

通过建立系统广义随机Petri 网模型，对BSO-MSS 进行性能评价，模拟结果显示无论增加存储对象（Storage Object，SO）还是客户端，系统性能都随之增加。并采用测试工具iozone 对系统原型与Lustre 系统作对比测试，测试结果表明写性能超过Lustre，读性能略比Lustre 好，并验证了BSO-MSS 的广义随机Petri 网模型。

首次将存储系统与元胞自动机相结合，利用元胞自动机的原理，解析BSO-MSS动力演变规律。构建了一个通用框架的BSO-MSSCA 概念模型框架，并在此基础上，分析了两种具体元胞自动机模型。基于存储对象负载分配模型是将SO 解析为元胞，模拟了一个简单的负载均衡分配的动态变化，高度概括了BSO-MSS 的演变过程。

基于数据对象访问行为模型则分析数据对象的访问频率对系统的影响，结合数据对象访问的特征和主动性，通过机械学习适当调整数据对象的访问行为频率，使系统朝着稳定方向发展。通过分析基于存储对象的负载分配模型和基于数据对象的访问行为模型的演变过程，可以看出系统具有主动性、共享性、并行性、相关性等特性，是一个自组织管理的对象存储系统。

大规模分布式存储系统中，元数据高性能服务、负载均衡以及扩展性已成为一个重要的研究热点。在元数据服务器中，将元数据分解为目录对象和文件对象，目录对象为定位性元数据，提供文件所在位置和访问控制；文件对象为描述性元数据，描述文件的数据特性。每一个元数据服务器（Metadata Server，MDS）负责所有目录对象和自身的文件对象，这样充分利用MDS 中Cache，提高Cache 的命中率，减少磁盘I/O 次数，而且能够动态扩展MDS。同时，以目录对象ID 和文件名为关键字的哈希值作为局部元数据查找表（Local Metadata Lookup Table，LMLT）的索引，获得相应的MDS_ID。一旦目录权限改变、更名、移动目录、修改权限等都不会造成元数据的迁移。通过Bloom Filter 算法将每个MDS 的LMLT 压缩成一个摘要，能够实现快速的元数据查找。同时采用主从备三重链式结构的MDS 服务，不仅在未提高硬件成本下能够保证系统高可靠性和可用性，而且根据热点访问进行迁移，实现负载均衡。SO 是BSO-MSS 重要组成单位，它与OSD 不同之处是本身具有“接口”与“状态”标识，由数据、属性和方法组成，这样对现有的T10 OSD 标准进行了扩充。由于数据对象是通常在一维空间中命名，传统文件系统管理大量数据对象的效率是极其低，采用线性哈希查找算法，由负载因子控制分裂和合并，与传统文件系统的树结构查找相比，哈希法查找时间复杂度为O(1)。同时，针对Ext2 文件系统中数据访问至少需两次以上的磁盘操作特性，将数据的块地址和长度链接在一起，作为对象的扩展属性，连同数据对象一起存储到磁盘中，这样无论数据对象大小为多少，磁盘访问次数仅为两次。在BSO-MSS 中，负载与众多因素相关，如请求队列长度、CPU 处理能力、内存大小、网络带宽、磁盘带宽和磁盘容量等。负载柔性放置策略不仅考虑网络的影响，而且考虑SO 之间存在差异，并设置权重，权重大的SO 担负较多的负载。依据SO 属性中信息统计出负载特征，以系统响应时间为代价，自适应选择SO 数目，采用不同大小的分条进行存储，使BSO-MSS 具有更高的性能、可扩展性和自适应负载均衡能力。