**睿普云存储管理系统**

**设计说明书**

**一、引言**

**目的**

随着互联网技术的飞速发展，各种网络工具的普及，越来越多的信息开始在网上以指数级别的速度飞速增长。然而这些海量的数据会给数据的存储管理带来越来越多问题，比如数据管理困难、数据存储成本高、可靠性低，索引比较困难等难题。于是，越来越多人和企业开始思考解决存储管理的诸多问题。而建立云存储服务管理系统成为解决海量数据的管理和存储问题以及这些数据的索引问题的关键。

本设计说明旨在为睿普云存储系统提供详细的架构设计、功能模块设计和接口设计，为进行下一步开发提供指导和说明。

**二、系统总体设计**

**2.1需求概括**

本软件采用传统的软件开发生命周期的方法，采用自顶向下，逐步求精的结构化的软件设计方法。

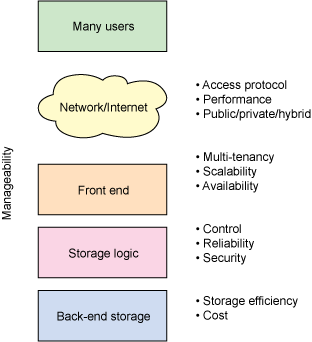
本软系统要有以下几方面的功能

1. 存储设备资源池添加节点
2. 存储设备资源池删除节点
3. 存储设备的挂载和卸载
4. 共享存储卷的挂载和卸载
5. 存储数据的展示
6. 存储数据的报表管理

**2.2通用架构**

云存储架构主要关乎以一个高度可扩展和多租户的方式按需交付存储。通用（参见 [图 1](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "fig1)）的云存储架构包含一个导出 API 以访问存储的前端。在传统的存储系统中，这个 API 是 SCSI 协议；但是在云环境中，这些协议在演化。在那里您可以找到 Web 服务前端、基于文件的前端，甚至更多传统前端（比如 Internet SCSI 或 iSCSI）。在前端后面是一个中间件层，我将它称作存储逻辑。该层通过传统的数据放置算法（考虑地理布局）实现各种功能，比如复制和数据简缩。最后，后端实现对数据的物理存储。这可能是一个实现特定功能的内部协议或物理磁盘的一个传统后端。

##### 图 1. 通用的云存储架构



从图 1 中，您可以看到当前云存储架构的一些特征。注意，没有特征在特定层中是独有的，而是充当本文探讨的特定主题的指导。这些特征的定义见 [表 1](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "table1)。

##### 表 1. 云存储特

| **特征** | **说明** |
| --- | --- |
| 可管理性 | 以最少的资源管理系统的能力 |
| 访问方法 | 公开云存储所用的协议 |
| 性能 | 根据宽带和延迟衡量的性能 |
| 多租户 | 支持多个用户（或承租者） |
| 可扩展性 | 通过扩展满足更高要求或以得体的方式加载的能力 |
| 数据可用性 | 对一个系统的正常运行时间的衡量 |
| 控制 | 控制系统的能力 — 特别是为成本、性能或其他特征进行配置 |
| 存储效率 | 度量如何高效使用原始存 |

### 可管理性

云存储的一个重点是成本。如果客户可以购买并在本地管理存储，而不是在云中租赁它，那么云存储市场就会消失。但是成本可划分为两个高级类别：物理存储生态系统本身的成本和管理它的成本。管理成本是隐式的，但却是总体成本的一个长期组成部分。为此，云存储必须能在很大程度上进行自我管理。引入新存储（其中系统通过自动自我配置来容纳它）的能力和在出现错误时查找和自我修复的能力很重要。在未来，诸如自主计算这样的概念将在云存储架构中起到关键的作用。

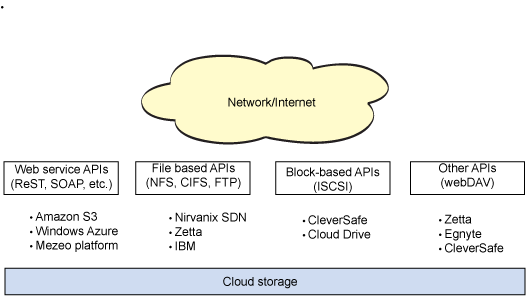
### 访问方法

云存储与传统存储之间最显著的差异之一是其访问方法（参见 [图 2](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "fig2)）。大部分提供商实现多个访问方法，但是 Web 服务 APIs 是常见的。许多 APIs 是基于 REST 原则实现的，即在 HTTP 之上开发（使用 HTTP 进行传输）的一种基于对象的方案。REST APIs 是无状态的，因此可以简单而有效地予以提供。许多云存储提供商实现 REST APIs，包括 Amazon Simple Storage Service (Amazon S3)、Windows Azure™ 和 Mezeo Cloud Storage Platform。

Web 服务 APIs 的一个问题是，它们需要与应用程序集成，以利用云存储。因此，对云存储也使用常见的访问方法来提供即时集成。例如，NFS/Common Internet File System (CIFS) 或 FTP 等基于文件的协议，iSCSI 等基于块的协议。Nirvanix、Zetta 和 Cleversafe 等云存储提供商提供这些访问方法。

尽管上面提到的协议是最常用的，但也有适合云存储的其他协议。最有趣的其中一个是基于 Web 的分布式创作与版本控制（WebDAV）。WebDAV 也基于 HTTP，且将 Web 作为一种可读写的资源加以启用。WebDAV 的提供商包括 Zetta 和 Cleversafe 等。

##### 图 2. 云存储访问方法



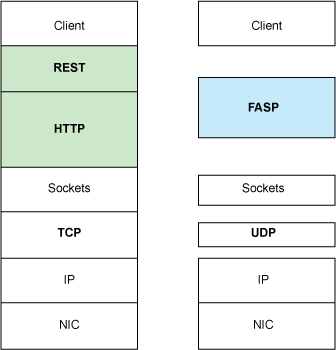
您还可以寻找支持多协议访问的解决方案。例如，IBM® Smart Business Storage Cloud 从同一存储虚拟化架构同时启用基于文件（NFS 和 CIFS）的协议和基于 SAN 的协议。

### 性能

性能表现为很多方面，但是在用户与远程云存储提供商之间移动数据的能力是云存储最大的挑战。问题就是 TCP，它同时也是互联网的主力。TCP 基于数据包确认从对等端点控制数据流。数据包丢失或延迟到达情况下将启用阻塞控制，进一步限制性能以避免更多全局网络问题。TCP 适用于通过全局 Internet 启用小量数据，但不适用于会增加往返时间（RTT）的大型数据移动。

通过 Aspera Software，Amazon 解决了这个问题，方法就是从程式中删除 TCP。且开发了一个称为 Fast and Secure Protocol (FASP™) 的新协议，以在大型 RTT 和严重数据包丢失情况下加速批量数据移动。关键是 UDP 的使用，它是 TCP 的缔约方传输协议。UDP 允许主机管理阻塞，将这个方面推进到 FASP 的应用层协议中（参见 [图 3](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "fig3)）。

##### 图 3. Aspera Software 上的 Fast and Secure Protocol



通过标准（非加速）NICs、FASP 有效使用应用程序可用带宽，并移除传统的批量数据传输模式的基本瓶颈。[参考资料](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "artrelatedtopics) 部分提供在传统 WAN、洲际传输和有损卫星链接中 FASP 性能相关的一些有趣统计信息。

### 多租户

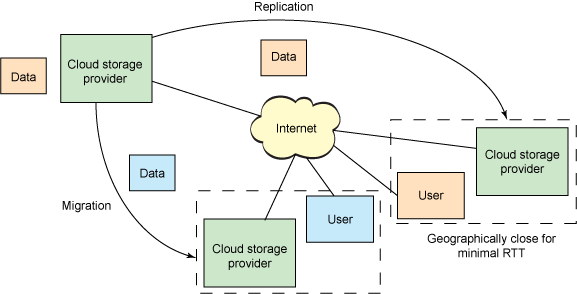
云存储架构的一个关键特征称为多租户。这只是表示存储由多个用户（或多个 “承租者”）使用。多租户应用于云存储堆栈的多个层，从应用层（其中存储名称空间在用户之间是隔离的）到存储层（其中可以为特定用户或用户类隔离物理存储）。多租户甚至适用于连接用户与存储的网络基础架构，向特定用户保证服务质量和优化带宽。

### 可扩展性

您可以从多个方面看待可扩展性，但正是云存储的随需视图使其最具吸引力。扩展存储需求（向上和向下）可改善用户成本，提高云存储提供商的复杂性。

不仅要为存储本身提供可扩展性（功能扩展），而且必须为存储带宽提供可扩展性（负载扩展）。云存储的另一个关键特性是数据的地理分布（地理可扩展性），支持经由一组云存储数据中心（通过迁移）使数据最接近于用户。对于只读数据，也可以进行复制和分布（使用内容传递网络完成）。这如 [图 4](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "fig4) 所示。

##### 图 4. 云存储的可扩展性



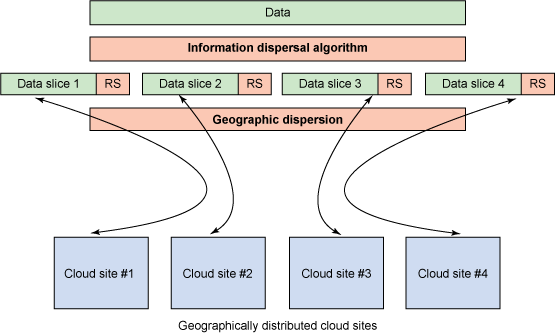
在内部，一个云存储架构必须能够扩展。服务器和存储必须能够在不影响用户的情况下重新调整大小。正如在 [可管理性](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "manageability) 部分所讨论的，自主计算是云存储架构所必需的。

### 可用性

如果一个云存储供应商有用户的数据，它必须能够应求将该数据提供给用户。鉴于网络中断、用户错误和其他情况，这很难以一种可靠而确定的方式予以提供。

有一些有趣而新颖的方案可用于解决可用性，比如信息传播。一家提供私有云存储的公司 Cleversafe（稍后介绍）使用 Information Dispersal Algorithm (IDA) 来在发生物理故障和网络中断的情况下实现更高的可用性。IDA 是由 Michael Rabin 最初为电信系统而创建的一种算法，它支持使用 Reed-Solomon 代码对数据进行切片处理，以便在数据丢失的情况下实现数据重建。此外，IDA 允许您配置数据切片的数量，这样一来，可以为一个可接纳故障将数据对象分割成 4 个切片，对 8 个可接纳故障分割成 20 个切片。与 RAID 类似，IDA 支持通过原始数据的子集重建数据，含有一定数量的代码错误开销（依赖于可接纳故障的数量）。这如 [图 5](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "fig5) 所示。

##### 图 5. Cleversafe 用于最大化数据可用性的方法



有了为数据切片的能力以及 cauchy Reed-Solomon 纠错码，就可以将切片分发到地理上分散的站点进行存储。对于大量切片（p）和大量可接纳故障（m），最终开销是 p/(p-m)。因此在 [图 5](https://www.ibm.com/developerworks/cn/cloud/library/cl-cloudstorage/index.html" \l "fig5) 中，p = 4 且 m = 1 的存储系统的开销是is 33%。

IDA 的缺点在于，它是处理密集型的，无硬件加速。复制是另一个有用的技术，且由各个云存储提供商实现。尽管复制技术引入了大量开销（100%），但可以简单而高效地提供它。

### 控制

一名客户控制和管理其数据存储方式及其相关成本的能力很重要。许多云存储提供商实施控制，使用户对其成本有更大的控制权。

Amazon 实现 Reduced Redundancy Storage (RRS)，为用户提供最小化总存储成本的一种方式。数据是在 Amazon S3 基础架构内复制的，但使用 RRS，数据复制次数较少，且存在丢失数据的可能性。这适用于可重新创建的或在其他地方有副本的数据。Nirvanix 还提供基于策略的复制来对如何以及在何处存储数据提供更细粒度的控制。

### 效率

存储效率是云存储基础架构的一个重要特征，特别是将重点放在总成本上。下一部分专门介绍成本，但是该特征更多地是关于对可用资源的高效使用，而非成本。

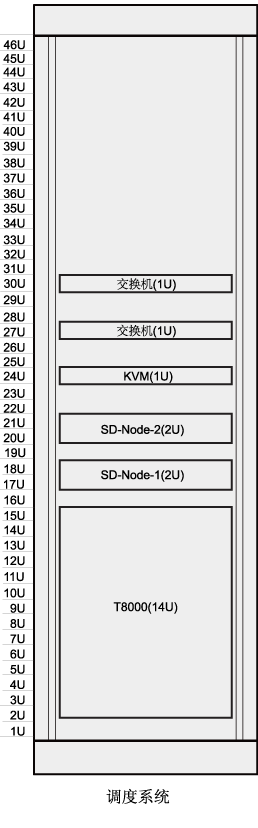
要使一个存储系统更高效，必须存储更多数据。一个常见的解决方案就是数据简缩，即通过减少源数据来降低物理空间需求。实现这一点的两种方法包括压缩 — 通过使用不同的表示编码数据来缩减数据 — 和重复数据删除 — 移除可能存在的相同的数据副本。虽然两种方法都有用，但压缩方法涉及到处理（重新编码数据进出基础架构），而重复数据删除方法涉及到计算数据签名以搜索副本。

## **2.3  系统配置**

介绍云存储的硬件外观和硬件配置，以及各硬件部件在整个系统中的作用，便于用户直观了解云存储的硬件结构和配置。体现云存储的大存储容量、高存储性能和良好的扩展性等特点。

云存储由1套调度系统和多套资源组组成。

调度系统的外观如[图2-2](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0007_fig01)所示。

**图2-2**  调度系统的外观  


调度系统的硬件组成如[表2-1](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0007_tab01)所示。

**表2-1**  调度系统的硬件组成

| **设备** | **型号** | **数量** | **主要功能** |
| --- | --- | --- | --- |
| 机柜 | 电信机柜 | 1 | 承载存储设备 |
| T8000 | T8223 | 1 | 配置6块刀片，负责云存储资源服务和调度管理 |
| SD-Node | T3200 | 2 | 负责系统数据存储 |
| 交换机 | S5352C-SI | 2 | 48口交换机，负责调度系统与资源组及用户网络通信 |

云存储配置多套资源组，一套资源组由资源服务系统和分布式文件系统组成，[图2-3](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0007_fig02)以第一套资源组为例说明其硬件组成。

**图2-3**  资源组的外观  


资源组的硬件组成如[表2-2](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0007_tab02)所示。

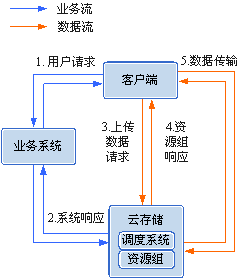
**表2-2**  资源组的硬件组成

| **区域** | **设备** | **型号** | **主要功能** |
| --- | --- | --- | --- |
| 资源服务系统 | 机柜 | 电信机柜 | 承载资源服务设备 |
| T8000 | T8223 | 配置12块刀片，负责云存储资源服务和云存储管理 |
| SD-Node | T3200 | 负责系统数据存储 |
| 交换机 | S5352C-SI | 48口交换机，负责资源服务系统内部网络通信、资源服务系统与分布式文件系统、资源服务系统与用户网络的通信 |
| MDS a | T3200 | 负责元数据存储和管理 |
| 分布式文件系统 | 机柜 | 46U电信机柜 | 承载存储设备 |
| MDS | T3200 | 负责元数据存储和管理 |
| OSN b | T3500 G2 | 负责对象存储和管理 |
| 交换机 | S5352C-SI | 48口交换机，负责资源服务系统内部网络通信、资源服务系统与分布式文件系统通信 |
| S5328C-SI | 24口交换机，负责分布式文件系统内部网络通信、分布式文件系统与资源服务系统通信 |
| a：Metadata Server，元数据服务器  b：Object Storage Node，对象存储节点 | | | |

## **2.4  系统组网**

介绍用户通过云存储实现数据读写的过程，以及云存储、资源组与用户网络的组网方式。便于用户了解云存储处理业务的基本流程，以及在整个网络中所处的位置和作用。

用户使用云存储进行数据读写的示意图如[图2-4](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0006_fig01)所示。

**图2-4**  用户业务读写示意图  


上图中各信息流的作用如下：

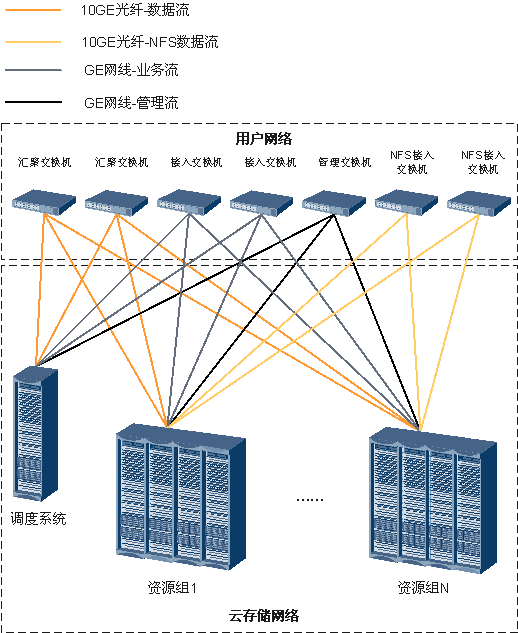
* 业务流

云存储与业务系统之间的数据流向，承载用户的业务请求流。

* 数据流

云存储与客户端之间的数据流向，承载用户的数据读写流。

云存储与用户网络的整体组网如[图2-5](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0006_fig02)所示。

**图2-5**  整体组网图  


上图中各组成部件的作用如下：

资源组

负责资源组的用户管理和认证、对象和容器管理、数据存储，并对外提供云存储服务接口。

调度系统

负责资源组的任务调度和资源分配。

汇聚交换机

负责云存储与客户端之间的网络连接，用于承载数据流。

接入交换机

负责云存储与业务系统之间的网络连接，用于承载业务流。

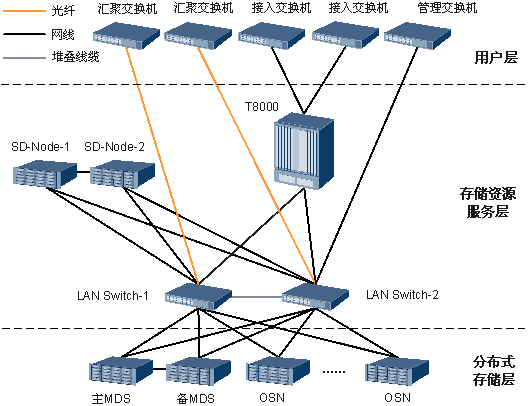
NFS接入交换机

负责外网与分布式文件系统之间的网络连接，用于承载NFS数据流。

管理交换机

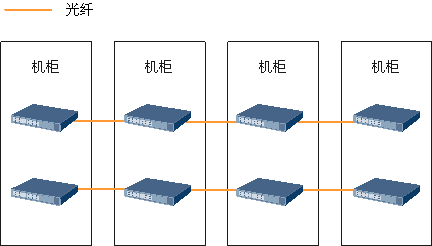
负责云存储与用户的管理网络之间的网络连接，用于承载设备管理流和系统管理流。

单套资源组与用户网络的组网如[图2-6](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0006_fig03)所示。

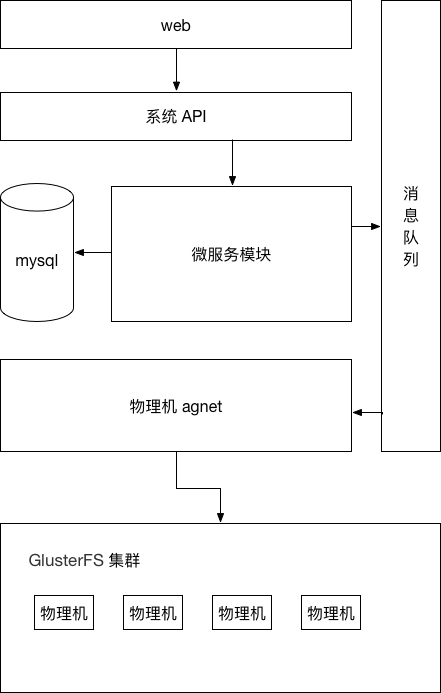
**图2-6**  单套资源组的组网图  


NFS接入交换机与每个资源组内的分布式文件系统扩展柜的交换机相连。

单套资源组内的机柜之间的交换机采用光纤直连的方式。网络连接拓扑图如[图2-7](https://support.huawei.com/view/contentview/getFileStream?mid=SUPE_DOC&viewNid=EDOC0000715804&nid=EDOC0000715804&partNo=j005&type=htm" \l "os_slu_des_0006_fig04)所示。

**图2-7**  资源组内部网络连接拓扑图  


## 2.4总体结构和模块接口设计



系统整体结构框架图

## 2.5模块功能逻辑关系

系统详细的模块信息所示：

系统详细的模块信息表

|  |  |
| --- | --- |
| **主模块** | |
| 模块名称 | 功能简述 |
| 块存储 | 管理块存储设备的创建挂载删除 |
| 共享券 | 管理共享卷的创建挂载删除 |
| Gluster 集群管理 | 管理 gluster 集群创建，节点删除修改操作 |
| Agent 命令管理 | 物理机 agent 接受命令操作 |

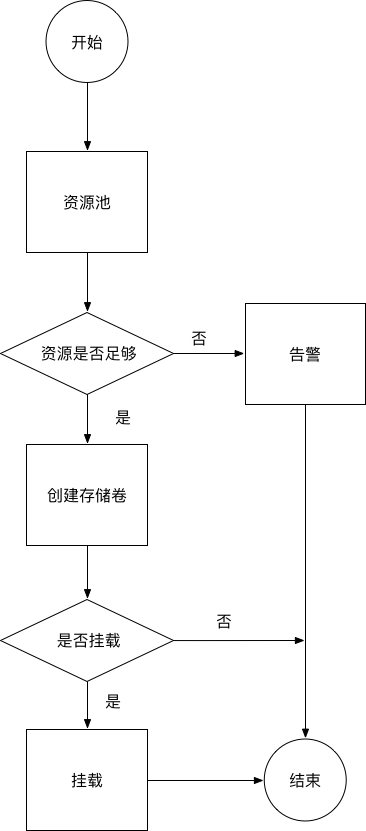
**2.6设计和描述**

存储系统主要负责管理睿普云系统的存储部分，包括块存储、共享卷、gluster 集群资源池管理，资源监控报警，日志统计，对外 API 接口等部分。块设备存储管理用户的块设备资源，用户可以创建块设备，在创建主机的时候，挂载该设备，也可以创建完成之后再进行挂载卸载等操作。共享券可以在用户的多个主机之间共享数据，记录用户的挂载关系，管理用户的共享存储数据。资源池管理用来管理底层的 gluster 的资源，这个功能是管理端进行资源管理的时候使用的，管理员可以管理通过集群管理创建新的节点，添加新的设备，对存储设备进行横向扩容。Agent 是部署在物理机上的服务端，用于底层设备管理，接受平台层命令执行底层操作。

**三、系统功能描述**

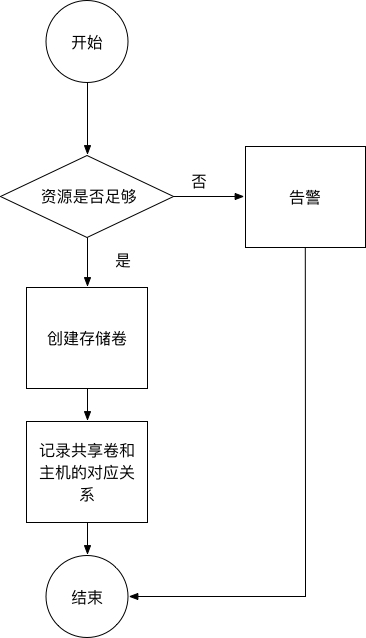
**3.1 块设备流程图和详细流程描述**

块存储设备在创建之前，gluster资源池中必须有足够的资源，当用户进行创建的时候，为用户创建对应大小的存储卷。当用户进行挂载操作的是，为用户进行对应的挂载操作，将用户创建的卷挂载到用户主机的对应挂载点上。

****

**3.2共享存储流程图和详细流程描述**

共享卷是创建，在物理机上为用户创建共享存储空间，当用户创建主机的时候进行挂载操作。共享卷无法进行动态挂载，需要用户创建主机的时候进行选择挂载。

****

**3.3资源池管理功能流程图和详细流程描述**

资源池管理，用于自动创建 gluster 集群，添加存储节点，创建和删除存储卷等操作。

# 

# 四、接口设计

## 4.1接口列表

GET /storage/findContainerStorage 查询容器的存储

POST /storage/updateStorage 修改存储名称

POST /storage/createStorageVhd 只创建镜像文件

POST /storage/delStorage 删除存储

POST /storage/updateStorageList 更新存储状态

POST /storage/formatImage 格式化镜像文件

POST /storage/formatImageParameter 格式化镜像文件回调

POST /storage/queryNoboundStorageContainer 查询没挂在存储的主机

POST /storage/updateMountpath 修改存储挂载点

POST /storage/createStorageVhdAssetCost 创建存储并支付接口

POST /storage/getCostPurchase 购买存储费用计算接口

POST /storage/getRenewalStorageCostPurchase 续费价格计算接口

GET /storage/findOutTimeStorage 查询过期块存储

POST /storage/storageManage 查询存储列表

POST /storage/deleteImage 删除镜像文件

POST /storage/deleteImageParameter 删除镜像文件回调

POST /storage/volume/updateMountpath 修改共享卷挂载点

POST /storage/volume/getCostPurchase 计算存储费用

POST /storage/volume/findVolumeStorageByUser 分页查询共享存储

POST /storage/volume/createVolumeStorage 创建共享存储

POST /storage/volume/deleteVolume 删除共享存储

POST /storage/volume/updateVolume 修改存储大小

POST /storage/volume/updateVolumeParameter 修改存储大小回调

POST /storage/volume/updateVolumeStatus 更新共享存储状态 type:0 恢复 10 彻底删除

GET /storage/volume/queryContainerVolume 查询这个容器的共享存储信息

POST /storage/volume/findVolumeStorageByUserAll 查询当前用户共享卷不分页

POST /storage/volume/finVolumeStorageByContainer 查询当主机下挂在的共享存储

POST /storage/volume/updateVolumeStorageName 修改存储名称

POST /storage/volume/deleteVolumeStorage 删除存储回调

POST /storage/volume/queryVolumeStaorageById 查询当前共享卷信息

POST /storage/volume/getStorageVolumenBindContainer 分页查询当前共享存储绑定的主机信息

POST /storage/volume/createStorageVolumenAssetCost 创建存储收费业务

POST /storage/volume/getStorageVolumenAssetRenewPrice 计算共享存储续费价格

POST /storage/volume/getExpansionCostPurchase 计算存储扩容价格

GET /storage/volume/findOutTimeVolume 查询过期共享存储

POST /storage/volume/isHasContact 查询共享存储是否被绑定