基于Linux-ZFS的分布式块存储系统设计

研究中心 朱韬 2013-01-31

# 云计算与块存储

云计算目前对于存储的需求日益增大，目前多家厂商和开源组织都发布了多种基于网络的分布式存储产品。此类产品大致分为两种：分布式对象存储，分布式块存储。其中分布式对象存储产品已经比较成熟，在采用云计算的互联网企业中早已开始大规模使用，主要是采用软件方式将信息分散存储于多台服务器上，相互实现备份，其短板是不能以系统文件方式存储信息，只能通过专门的软件来将信息对象化后进行存储。而分布式块存储设备，由于依赖操作系统底层功能，所以开发专门的产品难度很大，稳定性和可靠性都需要复杂而漫长的测试，因此目前在云计算的整体系统中没有单独研发成功的分布式块存储产品，大部分需求都是通过以往的成熟技术组合而成的块存储服务来提供。

成熟的分布式块存储解决方案，大都是基于光纤存储网络实现，性能和可靠性都有保障，但是成本高，维护费用昂贵，部署上也不太灵活，不适合中小规模企业采用。

我们的目的是采用一般相对廉价的设备，依靠原有的成熟技术实现满足需求的分布式块存储系统，保证存储的一定性能和可靠，同时做到相对简单和方便的扩展性和可维护性。

# 分布式块存储需要解决的问题

1. 实现数据的分布式存储
2. 将多个独立服务器上的本地存储资源集中起来统一管理和分配
3. 对数据做到备份、容灾、快速恢复
4. 动态容量扩展
5. 不停止服务的状态下更换硬件设备
6. 对外提供块存储设备
7. 对外提供共享存储服务

# ISCSI技术在分布式块存储中的价值

ISCSI技术是通过网络将一台主机上的存储设备导出给另外一台主机，既然是存储设备的导出，则说明ISCSI完全是系统底层提供的功能，与上层应用之间没有关联，因此性能上有所优势。但由于ISCSI是设备协议，因此它只能导出存储设备，无法管理存储内容和设备上的文件系统，所以ISCSI技术主要面向针对网络中各主机或虚拟机的模拟本地存储。

# ZFS在分布式块存储中的价值

ZFS是个文件系统，最早由SUN公司研发，目前FreeBSD和Linux都有支持的版本。ZFS是采用全新的文件系统理念，将存储设备和存储模式忽略（用户角度），极大地提高了操作系统对于存储的扩展性和维护。它将传统的存储标准集成在一起，如RAID技术和虚拟逻辑卷等，使用户方便的实现应用中绝大部分的存储需求。

## ZFS应用于分布式块存储中的功能特点

1. 整合多个/多种存储硬件设备
2. 灵活设置子文件系统
3. 灵活设置文件系统挂载点
4. 动态生成虚拟存储设备卷
5. 动态调整存储容量
6. 动态扩容
7. 生成文件系统快照
8. 方便恢复、回退和重置文件系统
9. 有完整可靠的数据保护方案

# 基于Linux-ZFS的分布式块存储系统设计原理

我们采用Linux系统作为存储节点和块存储服务节点的基础操作系统，Linux应用广泛稳定且有丰富的技术支持资源，便于维护和升级。

采用zfsonlinux作为系统存储服务的基本工具，该软件包提供了Linux系统下关于zfs文件系统的大部分功能，需要编译后将zfs文件系统模块加载如Linux系统之中。

采用iscsi技术实现存储节点与块存储服务节点间的连接，存储节点将本地存储设备以iscsi-target模式导入块存储服务节点，服务节点以iscsi-initiator的方式将存储节点导出的设备加载入服务节点本地存储设备树中。最后由zfs文件系统统一整合。

采用iscsi-target技术向客户端提供块存储设备，采用nfs和samba服务向客户端提供共享存储服务。

## 基本原理



### 存储节点设计原理

存储节点安装Linux操作系统作为基本系统平台，将本地存储设备以软件raid的方式统一整合，将整合后的完整raid存储设备作为iscsi的target端向服务器导出。

采用软件raid处于如下考虑：

1. 成本低
2. 实现简单
3. 不依赖于专用硬件
4. Raid阵列扩容、恢复等维护操作灵活

虽然软件raid在部分性能上低于基于硬件raid卡的阵列，但是软件raid相对硬件raid更加安全，硬件raid如果控制器设备出现故障，raid将很难恢复，而软件raid没有控制器硬件的辅助，一切功能都有软件实现，更加具有容灾性。

目前许多大型服务器厂商已经陆续抛弃了硬件raid而采用软件raid来实现存储设计，在系统运行性能大幅提升的趋势下，软件raid的性能也在逐渐与硬件raid持平，甚至某些性能还要高于硬件raid。

### 块存储服务器设计原理

块存储服务器节点安装Linux操作系统作为基础系统平台，编译zfsonlinux软件包，使得系统支持zfs功能。

将存储节点导出的iscsi设备，导入块存储服务器本地的存储设备树中，再由zfs将多个iscsi存储设备统一整合为一个zfs存储池（zpool）。

在zfs存储池中建立虚拟设备卷和子文件系统，将虚拟设备卷以iscsi-target的方式导出给客户端服务器或虚拟机使用，将子文件系统设置成nfs或samba服务供集群各服务器系统使用。

ZFS文件系统在分布式块存储服务器中的主要功能：

1. 整合多个/多种存储设备
2. 建立虚拟设备卷
3. 建立子文件系统
4. 动态调整个存储但愿的容量
5. 动态加入新存储设备
6. 动态移除、更换检测存储设备
7. 制作文件系统快照
8. 整个存储环境的备份和恢复

## 服务器硬件部署原理

分布式块存储系统采用集中管理分布存储的架构，由一台高性能服务器作为管理和提供服务的中心，由多台普通服务器作为数据存储的分布式存储节点，这样的结构便于维护管理，也便于容灾和扩容。

### 基本原理



### 网络性能要求

服务器节点的两个网口必须是高性能的，可以采用10G以太网卡，也可以将多个千兆以太网卡绑定为一个网口实现。服务器节点和交换机的连接也必须是高性能的，如果服务器采用10G以太网卡，则交换机上必须要有相应的10G以太网口，如果服务器采用多网卡绑定模式，对应于交换机上的网口也必须绑定。

存储节点和应用服务器均可采用千兆以太网设备。

### 服务器性能要求

服务器节点由于要采用zfs文件系统，所以必须是大内存的高性能服务器，而且网络设备也必须是高性能的，因为所有服务的提供和存储节点的连接都是通过网络接入到服务器节点。

由于zfs文件系统必须在本地保留一定的配置文件和相应的存储设备拓扑快照，因此服务器节点的本地存储（安装操作系统的存储设备）必须能够保证数据的安全（RAID实现）。

## 应用模式



### 块设备存储服务提供

ZFS可以在存储池中建立虚拟设备卷，并在系统/dev目录下生成相应的设备文件。虚拟设备卷的容量可以在创建时设置，也可以在后期扩容或缩小，将创建出的虚拟设备卷以iscsi-target的方式导出，就可以让需要添加类似本地存储的应用服务器或虚拟机加载使用，使用过程和个服务器本地存储设备相同。

同时可以根据块存储服务器上的iscsi服务配置将虚拟卷和应用服务器惊醒对应绑定，这样提高了运行的稳定性也提高了数据安全性，实现了不同应用间的数据隔离。

### 共享存储服务提供

ZFS可以在存储池中建立子文件系统，这些子文件系统的大小可以初始设定也可以后期更改，通过操作系统的服务设置，就可以将这些子文件系统作为NFS和SAMBA服务的提供者，供应用服务器或虚拟机挂载访问，实现共享存储服务。

# 维护保障模式

服务器节点本身系统数据安全由服务器本地的RAID机制完成，维护途径也是本地维护的方式，存储节点导出的iscsi设备被zfs文件系统统一管理，维护途径有zfs文件系统统一负责，存储节点本地的存储设备有个存储节点自身的raid机制实现各自管理维护。

服务节点的本地raid保证服务节点本身系统的数据安全；ZFS文件系统保证各存储节点间的数据安全；存储节点自身的raid保证存储节点本地的数据安全。

这样一来，各级存储服务都有相应的技术和保障途径，有机的实现了整体的数据安全和系统维护。

# 后续重要辅助功能

1. 系统中各服务器状态监控功能
2. 系统中各存储设备故障报警功能
3. 分布式块存储服务日志统计功能

# 相关技术参考

1. ZFS文件系统手册
2. ISCSI技术手册
3. Linux软件Raid技术手册
4. NFS服务应用手册
5. SAMBA服务应用手册
6. Linux网络设备配置手册