Panasas

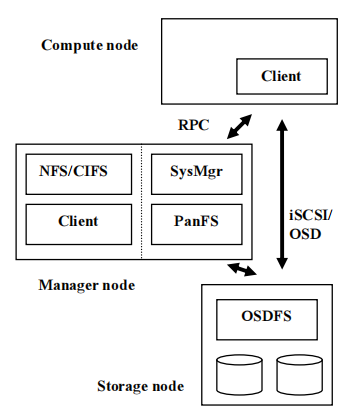
概述

Panasas公司于1999年由卡内基梅隆大学的Garth Gibson等人创建。其提供的并行文件系统PanFS是通用的并行文件系统，目前其主要的应用领域和luster类似，都是用于高性能计算（HPC）领域，这就要求其具有高可用，可扩展行，高性能，通过分布式锁提供的强一致性。

解决的问题

Panasas的两个总体主题是对象存储，它影响文件系统如何管理其数据，以及组件的集群，这允许系统在性能和容量上进行缩放。

Panasas存储系统是被设计为一个具有极强的可伸缩性的存储集群。同时，它也是一个生产系统，为世界上一些最大的计算集群、科学实验室、地震数据处理、数字动画工作室、计算流体力学等提供文件服务。在这些环境中，数百或数千个文件系统客户端共享数据，并在文件系统上生成非常高的聚合I/O负载。而 Panasas就为了支持数千个客户设计的、存储容量超过1兆字节的分布式存储系统。

设计架构

OSD:基于对象的存储设备， 是一种智能的存储节点。它将对象存储在本地OSDFS文件系统中，这也是当今磁盘驱动器的一种更智能的演变。它可以布局，管理和提供对象。OSD协议提供面向字节的数据访问、属性操作、对象的创建和删除以及其他几个专门操作。Panasas使用iSCSI传输来携带OSD命令。

Client:客户机（Client）是运行在Linux内核中的可安装内核模块。 内核模块实现了标准的VFS接口，使客户端主机能够挂载文件系统并使用存储系统的接口。

图 1 panasas的组成部件

OSDFS:OSDFS是存储节点使用实现对象存储原语的专用本地文件系统。 它们实现iSCSI目标和OSD命令集。 OSDFS对象存储和iSCSI target/OSD命令处理器是内核模块。 OSDFS关注传统的块级文件系统问题，如有效的磁盘臂、媒体管理（即错误处理）、高吞吐量以及OSD接口。

SysMerg：集群管理器(SysMerg)维护全局配置，并控制存储集群中的其他服务和节点。 有一个关联的管理应用程序，它同时提供命令行接口(CLI)和HTML接口(GUI)。 这些都是运行在管理器节点子集上的用户级应用程序。 集群管理器涉及存储集群中的成员资格、故障检测、配置管理和软件升级和系统重新启动等操作的总体控制。

Pan FS：元数据管理器(Pan FS)实现文件系统语义，并管理跨对象存储设备的数据条带。 这是一个运行在每个管理器节点上的用户级应用程序。 元数据管理器涉及分布式文件系统问题，如安全的多用户访问、保持一致的文件和对象级元数据、客户端缓存一致性以及从客户端、存储节点和元数据服务器崩溃中恢复。 容错基于本地事务日志，该日志被复制到不同管理器节点上的备份。

NFS/CIFS：NFS和CIFS服务为不能使用Linux可安装文件系统客户端的主机提供对文件系统的访问。 NFS服务是运行在内核中的标准FreeBSD的 NFS服务器的调优版本。 CIFS服务基于Samba并在用户级别运行。 反过来，这些服务使用本地实例文件系统客户端，它们运行在FreeBSD系统的内核中。 这些网关服务运行在每个管理器节点上，提供集群NFS和CIFS服务。

关键技术

对象存储：

OSD（基于对象的存储设备）与分布式存储的耦合使存储系统可以达到新的可伸缩性水平。这是在CMU的NASD项目中提出的承诺，现在已由Panasas和Luster存储系统证明，目前两者都提供10 Gb / s以上的速度。的存储带宽（总计超过100个客户端），并且有很大的扩展空间。

有效的OSD实施对于实现可扩展的高带宽至关重要。特别是，Panasas OSD文件系统（OSDFS）致力于最大程度地提高读写流量模式的磁头利用率。与基于块的接口相比，OSDFS的对象接口向存储设备公开的信息要多得多，从而使OSDFS可以执行更为强大的管理算法。OSDFS强调每个对象的有效布局，支持在高负载下最大化总吞吐量的支持，提供有效头调度的智能预取算法以及在许多并发I / O流之间平衡缓存利用率的缓存管理算法。

简化的管理模型：

Panasas存储系统将自己呈现为具有POSIX接口的文件系统，并隐藏存储管理的大部分复杂性。 客户端对整个系统有一个安装点。 /etc/fstab文件引用集群管理器(SysMerg)，客户端从中学习元数据服务实例的位置。 管理员可以在系统上线时添加存储，自动发现新资源。

自动容量平衡：

添加新的空存储节点、合并两个存储节点以及在故障后替换存储节点时，会发生容量不平衡。在后一种场景中，不平衡是我们的RAID重建的结果，它在每个存储节点上使用备用容量，而不是在重建过程中指定特定的。在替换失败的存储节点后导致系统有一个新的空存储节点。

Panasas是自动使用两种机制：被动平衡和主动平衡。被动平衡改变了存储节点根据其可用容量用于文件新组件的概率。当文件被创建，当它们的条纹大小被增加以包含更多的存储节点时，这将生效。主动平衡是通过将现有组件对象从一个存储节点移动到另一个存储节点，并更新受影响文件的存储映射来完成的。在传输过程中，文件被存储管理层透明地标记为只读，容量平衡器跳过正在积极写入的文件。因此，容量平衡对文件系统客户端是透明的。

总结

Panasas并行文件系统的设计说明了其可扩展性。该设计使用运行OSDFS对象存储的存储节点，以及运行文件系统元数据管理器、集群管理器和可以通过NFS和CIFS重新导出的Panasas文件系统客户端的管理器节点。可扩展性来源于每个存储节点的均衡性，包括磁盘，CPU，内存和网络带宽资源。 Panasas存储集群是用于块分配的并行机器，并且它是进行RAID重建的并行机器。因为每个存储节点都私有地管理自己的驱动器，而且每个管理器都参与了扩展到存储集群的解密奇偶校验组的重建。