PC 集群存储系统应用分析

Application Analysis On Storage Systems In PC Cluster Systems

贾亚军　张 峰　塔依尔　许 涛
（新疆油田分公司勘探开发研究院地球物理研究所　新疆乌鲁木齐　830013）

摘要：随着 PC 集群计算机技术的日益成熟发展，与 PC 集群并存的存储环境无法满足大规模并行计算环境的 I/O 需要。本文通过一系列的实验对比，提供了一种解决 PC 集群计算环境和存储环境不匹配的方法。

关键词：计算机应用　PC 集群　存储系统　蓝鲸文件系统

1　前言

随着石油地震勘探工作的不断深入，地震数据处理工作集中在寻找岩性圈闭、提高复杂构造成像效果等方面。有效的解决办法之一是将三维地震处理叠前时间/深度偏移技术纳入日常处理流程中。类似克希霍夫三维叠前偏移算法技术是一项消耗计算资源巨大的技术，使用常规的 UNIX 并行计算机技术所占用的资源是大多数企业无法承担的。因此越来越多的企业将基于 Linux 系统的 PC 集群技术引入高性能计算领域解决实际生产问题，然而与之相对应的是 PC 集群存储解决方案发展相对滞后，PC 集群结构中存储环境的效率很难与大规模计算环境的效率很好的匹配。因此各企业在构建 PC 集群系统时都根据企业不同业务特点和需求，采用不同方案构建相应的存储系统。

2　蓝鲸文件系统对 NFS 存储方案的改造

目前，为解决 PC 集群存储环境和计算环境不匹配问题，各研究机构和应用企业从许多方面提出了大量的解决方案。根据这些方案的实现原理，它们可以分为两类：

第一类方案可归结为“硬件扩容”方案，这些方案包括构建 SAN 存储网络、利用 10 千兆以太网络搭建新的存储网络、将 PC 集群技术引入存储环境中等。这些方案主要特点就是尽量提高存储环境的存储带宽，从而尽量提高存储设备的使用率。

另一类方案则倾向于“软件改造”，采用这一思想的设计方案都认为，制约存储环境效率发挥的主要障碍在于现有的面向网络存储应用的协议（NFS 协议）开销过大且效率较低，不能满足于大规模节点同时对网络存储设备进行访问的需求，因此有必要开发一种新的适应这种需求的网络存储协议。这些方案包括：GPFS、DAFS、PVFS 和蓝鲸网络并行文件系统（BWFS）等。

分析以上各种方案的实现原理，笔者认为进行软件改造的方法是目前解决问题的最为可行的方法，即通过某种方法改进 NFS 网络文件协议，从而使存储系统能够充分发挥现有网络带宽，提高存储效率。比较各种相关方案后，笔者认为蓝鲸文件系统（BWFS）是目前对使用 NAS 构建的 PC 集群存储系统进行改造较为适合的方案之一。

蓝鲸网络存储系统（BWFS）是目前在国内网络存储技术研究领域处于领先地位的一种网络分布式并行文件系统。该系统的核心就是开发一套新型的网络文件协议（BWFS）来替代 NFS 网络协议。通过分析发现，NFS 协议效率较低的主要原因是文件访问操作和数据读写操作都由文件服务器执行，需要完成的指令较多，操作复杂从而影响了存储系统整体读写效率，存储带宽无法充分发挥效率。蓝鲸网络存储系统通过将网络文件访问控制指令和数据读写操作指令分离由不同的处理器进行处理，元数据控制器只执行文件访问控制指令，而数据读写操作无需通过元数据控制器而是直接访问存储资源，从而有效地提高了网络带宽使用率，可以达到 90% 以上的有效负荷网络使用效率。由于蓝鲸存储系统网络文件协议在语义上与 NFS 网络文件协议保持一致，因此采用该项技术 PC 集群结构不需

要进行大的改动，应用软件无需进行修改即可透明访问相应的存储资源；另一方面，通过适当改造，部分NAS存储设备能够融入蓝鲸存储系统的管理中。蓝鲸存储系统目前已应用于国防工程等多个领域，然而在石油地震勘探领域尚无应用先例。为了将蓝鲸存储系统合理融合到石油地震勘探数据处理环境中，新疆油田分公司与中科院国家高性能计算机工程技术中心进行合作，在PC集群环境中进行了测试和部署并获得了令人满意的结果。下面就对此次测试进行说明。

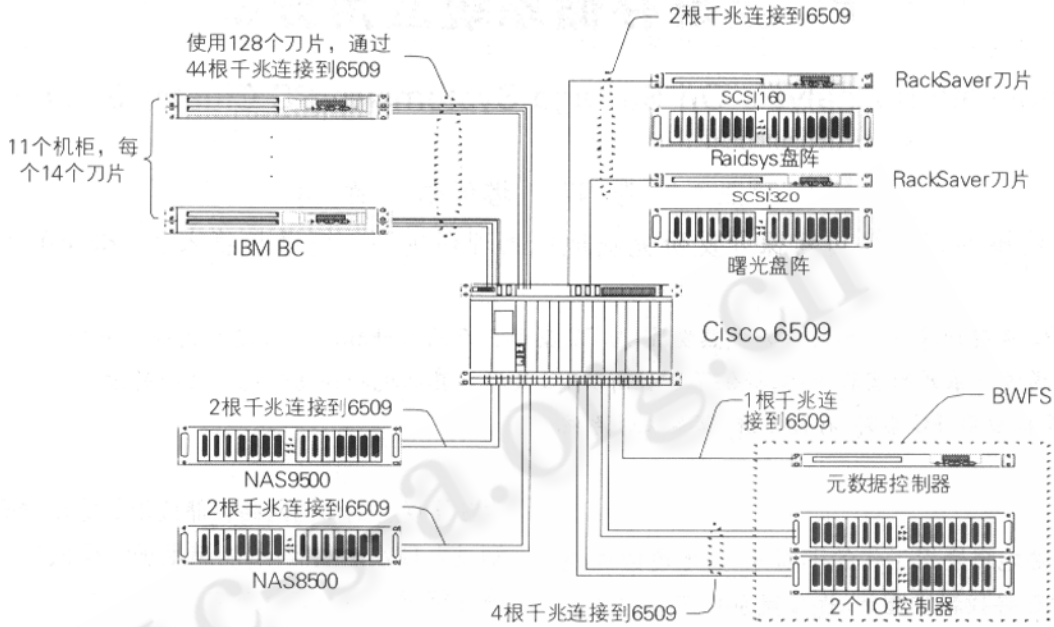

图 1　测试系统网络拓扑结构

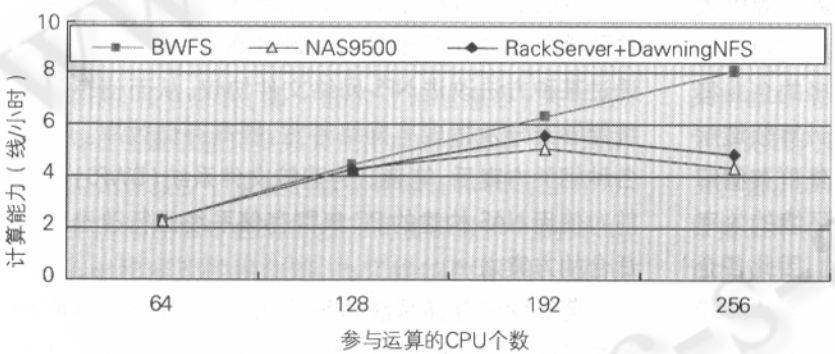

图 2　BWFS、NAS 和 DAS 的工作效率对比图

测试系统的网络拓扑结构如图1所示，测试主要集中在NAS、直连盘阵和BWFS之间实际工作环境中性能比较，该测试的核心交换机是Cisco公司生产的6509网络交换机，该交换机提供高达720Gb/S的背板交换速度，充分满足PC集群数据交换速度需求。

通过对比，在石油地震勘探叠前偏移处理工作中，随着PC集群计算资源规模的增加，使用BWFS存储技术能够更好地保证计算资源效率的线性增长。当参与运算的计算节点CPU个数达到一定规模后（256CPU），使用NAS存储资源和直连盘阵资源构建的存储系统已经无法满足计算能力的增强从而拖累计算环境使之无法发挥效率，甚至还造成了计算效率大大降低。而使用BWFS计算效率依然保持线性关系，如图2所示。

对于石油地震勘探常规处理流程中存储资源I/O要求较高的格式转换模块测试中，依然能够看到BWFS的优越表现，如图3所示，可以充分满足相关作业的工作需求，并且在同时有15个格式转换作业同时执行响应速度依然保持线性增长的关系，其工作效率的优越性不言而喻。

通过以上测试，可以认定蓝鲸文件系统（BWFS）比较适合PC集群石油勘探处理计算环境中海量存储和计算规模扩展性要求的系统应用。　（下转第 111 页）

(上接第 108 页)

3 结论

通过大量分析,可以认为制约存储系统发挥效率的根本原因是“网络存储协议 NFS 协议开销大,效率低,不能适应 PC 集群大规模计算的应用现实”,因此,从准葛尔盆地实际勘探数据处理情况出发,采用蓝鲸文件系统存储技术构建 PC 集群的存储体系有效地解决了存储环境的效率与计算环境的效率不匹配的问题。同时该构建方案还具备较高的性能价格比、易维护以及结构简单等特点,利用蓝鲸系统或相似性其他技术改造目前 PC 集群系统的存储体系将是一种快速提高 PC 集群使用效率的最佳方案。

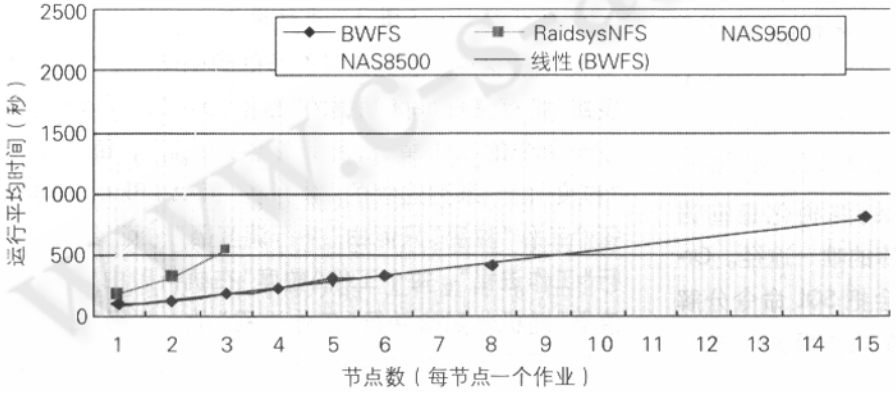

图 3 常规格式转换计算测试结果比较图

参考文献

1 中国科学院国家高性能计算机工程技术研究中心,《蓝鲸 1000 网络存储系统技术白皮书》.