新疆天文台 NSRT 观测数据存储系统 *

张海龙 [1,2]，朱艳 [1,3]，聂俊 [1,2]，袁建平 [1]，吴刚 [1]，刘俊 [1]，王杰 [1]，王万琼 [1]，冶鑫晨 [1]，托乎提努尔 [1]，张萌 [1,3]

（1. 中国科学院新疆天文台，新疆 乌鲁木齐，830011；2. 中国科学院射电天文重点实验室，江苏 南京，210008；3.中国科学院大学，北京，100049）

摘要：新疆天文台南山 26m 射电望远镜（简称 NSRT）经过多年观测积累了大量的科学数据。针对 NSRT 天文观测数据的在线存储与备份问题，建设了远程、异地、容灾备份系统，在新疆天文台本部及南山观测站分别建设了可独立运行的存储系统，实现了两套存储系统间的远程、异地数据实时容灾备份。以基于对象的存储技术 Lustre 为基础实现了存储系统，并对存储的读写性能进行了详细测试。建设的容灾备份系统有效解决了新疆天文台观测数据及次生数据的在线存储与数据安全问题。

关键字:观测数据；存储；备份；虚拟天文台；数据安全；

中图分类号：P111.5; TP391 文献标识码:A 文章编号:1672-7673(2018)

---

* 基金项目：国家自然科学基金(U1531125);国家重点基础研究发展计划（973）项目（2015CB857100）;中国科学院青年创新促进会;中国科学院天文台站设备更新及重大仪器设备运行专项经费资助.

收稿日期:2017-09-06;修订日期:2017-09-28

作者简介：张海龙，男，博士. 研究方向：数据密集型研究.Email：zhanghailong@xao.ac.cn

SKA 的先驱阵列望远镜 MWA[1]，由 2048 面低频阵列望远镜组成，相关后每秒归档数据在 400MB 左右，数据首先在线归档在 MRO 天文台的数据存储中，然后通过 10Gpbs 专线将数据实时传输备份到位于 MRO 700 千米以外的 Pawsey 数据中心，同时 Pawsey 数据中心数据按需求再通过 1Gpbs 线路备份到 MIT,USA、VUW，New Zealand、RRI，India[1]。

中国科学院国家天文台数据中心[2]是中国目前最大的天文数据库，包括国家天文台下属的各天文观测设备产生的天文数据，还有部分其它天文台站的观测数据，目前数据中心部分数据备份在中国科学院网络中心，部分数据备份在阿里云平台。

中国科学院紫金山天文台对外开放的数据库[3]包括毫米波射电天文数据库、太阳射电频谱观测数据库、近地天体望远镜数据库、太阳光谱数据库等，各数据库已实现在线访问，并建立了相应数据备份系统。

斯特拉斯堡天文数据中心[4]、欧洲南方天文台科学数据中心[5]、CSIRO ATNF 数据归档中心[6]、中国科学院上海天文台[7]、中国科学院云南天文台[8]等天文研究机构都分别建设了数据管理系统。

1、NSRT 数据情况简介

新疆 25m 射电望远镜[9]建成于 1993 年 12 月并投入使用，经过升级改造后口径扩大到 26m，新的 26m 射电望远镜简称 NSRT（NanShan Radio Telescope）。NSRT 承担着重要的国际合作及国内重大课题的天文观测任务，目前是欧洲甚长基线干涉网（EVN），国际动力测地网（IVS），俄罗斯低频 VLBI 网（LFVN），东亚 VLBI[10]网 4 个国际合作组织的正式成员。参加了 11 项国际合作计划，承担着国家攀登计划、大科学工程、绕月工程、火星探测、国家自然科学基金课题、中国科学院基础研究重点项目以及多项单天线国际合作天文观测研究任务和项目。

NSRT 开展了脉冲星、分子谱线、IDV 巡天和监测等多项课题，支持了银道面

[1] http://www.mwatelescope.org/telescope/data-archive
[2] http://www.china-vo.org/
[3] http://www.pmo.ac.cn/qt/twsjk/
[4] http://cdsweb.u-strasbg.fr/
[5] http://archive.eso.org/cms.html
[6] http://www.atnf.csiro.au/observers/data/index.html
[7] http://119.78.226.68/
[8] http://fso.ynao.ac.cn/dataarchive_ql.aspx
[9] http://www.xao.ac.cn/jgsz/ywtz/nsjd/25msd/
[10] https://en.wikipedia.org/wiki/Very-long-baseline_interferometry

磁场巡天、木星研究等观测。设备运行有效观测时间连创国内同类射电望远镜新高，在国内外天文观测研究中发挥了积极的作用。随着观测数据的猛烈增长，如何永久保存这些珍贵的天文观测数据，如何合理有效地解决这些数据的在线存储管理问题，如何高效地实现远程、异地容灾备份是新疆天文台 26 米射电望远镜运行中面临的一个重要课题[2]。

2000 年 1 月至 2002 年 6 月，NSRT 脉冲星观测系统由一个双通道室温接收机，带宽 320 MHz，中心频率 1540 MHz，消色散系统采用 2x128x2.5 MHz 模拟滤波器组实现，得到的脉冲星数据格式为“Timer”[3]。2002 年下半年低温接收机系统投入使用，制冷后的接收系统使天线灵敏度达到了 0.5 mJy[4]。2010 年 1 月，DFB(数字滤波器系统)投入使用，DFB 系统具有更高的时间分辨率，使得 NSRT 可以常规的观测到大约 280 颗脉冲星，其中包括 10 颗毫秒脉冲星，DFB 系统的数据记录格式为“Psrfit”，“psrchive”程序可以读取和分析数据。通过十多年的观测，脉冲星相关观测积累了大量数据, 目前已发布 56000 多条有效原始数据记录, 原始数据及处理后数据总量近 20TB[5]。

利用 NSRT 开展了分子谱线 OH, H2CO, NH3, H2O 等观测，从 2010 年开始，数字消色散系统应用后，产生的原始数据格式为 RPFits，获得的分子谱线原始数据经过校准之后，可用来估算星际介质，分子云的物理化学性质[6]，目前分子谱线相关已归档数据量在 5TB 左右。

自 2004 年起，利用 NSRT 的 6 厘米连续谱观测系统开展了河外射电源的流量监测，包括北天 blazar 天体的大样本快速光变巡天，以及 Fermi AGN 的长期射电流量监测等观测项目[7]。连续谱观测系统终端由马普射电所研发的便携式终端实现，其工作的中心频率为 4800MHz，带宽为 600MHz。原始数据为 FITS 格式，观测数据需要进行指向、大气不透明度、增益以及时间依赖等校准，最终转换成射电源绝对流量后可应用于科学研究[8]。经过多年的观测和积累，连续谱观测获取了～800 个射电源的共计约 250000 条有效原始数据记录，数据量约 10TB。

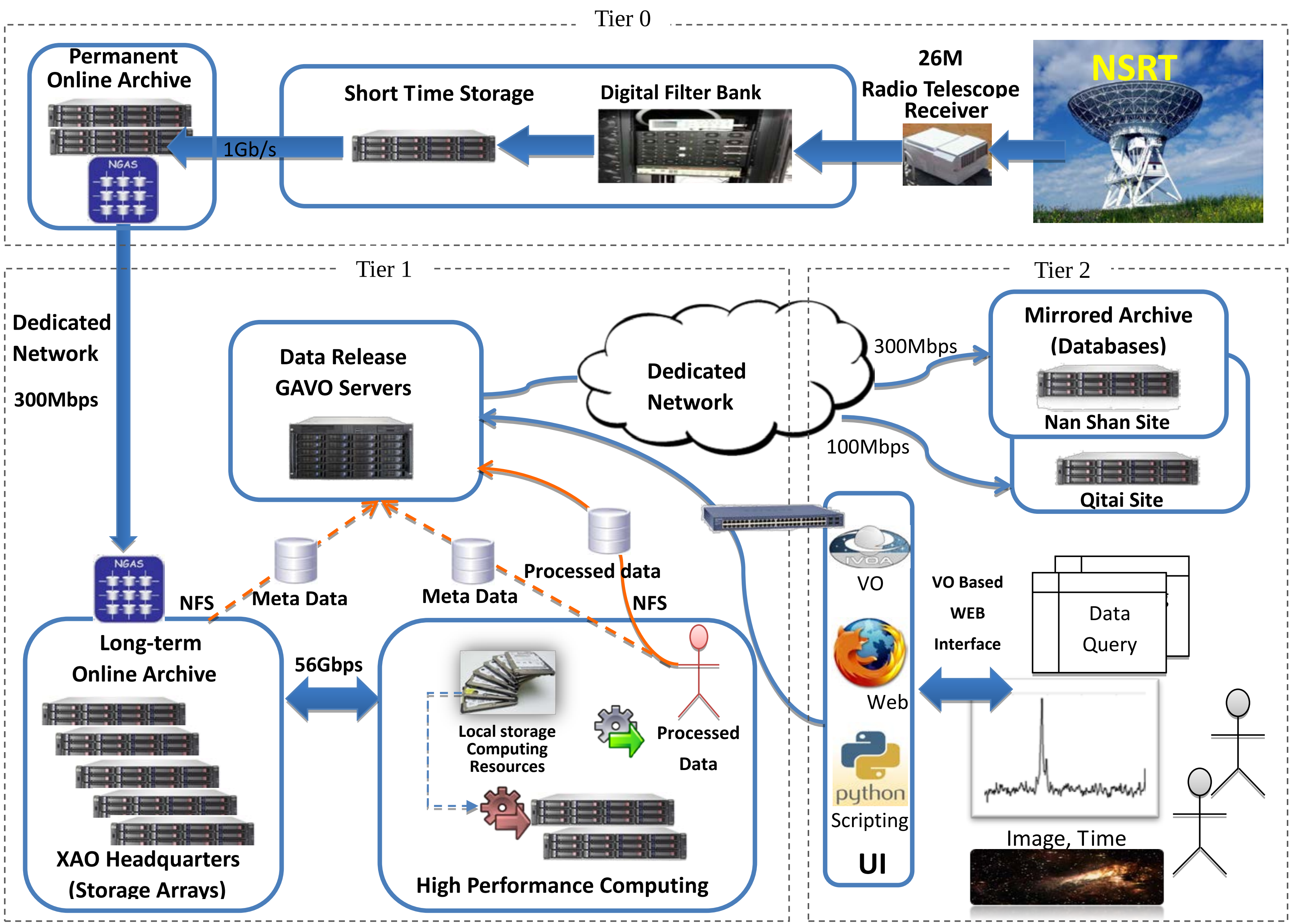

图 1 新疆天文台 26 米射电望远镜数据存储备份系统设计

2、数据存储与备份系统设计

NSRT 数据存储备份系统设计如图 1，设计共包含三个部分，第一部分实现观测数据获取与在线归档，第二部分完成原始观测数据的异地备份，第三部分负责数据发布共享。

第一部分：数据获取与在线归档部分在南山观测基地完成，由 26 米望远镜、接收机系统、DFB 系统、数据暂存系统及归档系统组成。数据获取以脉冲星观测为例，脉冲星的数据采集和处理以及数据记录由望远镜接收系统完成，数据采集程序实时完成数据采集、数据预处理、脉冲星周期计算和周期叠加、数据存盘、消色散数据显示、观测纲要查询、图形输出、天线姿态控制等几个任务。观测中典型的采样速率为 1 ms，每次观测时间由脉冲星在该观测频段的流量强度决定，一般为 2—16 分钟。数据经过预处理后写入暂存服务器中，经过科学家确认有效后数据将实现永久归档。

第二部分：原始观测数据的异地备份（新疆天文台本部位于南山观测基地北部 100KM 左右）通过南山观测基地到新疆天文台本部间的专线实现，专线速度 300Mbps, 数据由南山的 Permanent Online Archive 同步到新疆天文台本部 Long-term Online Archive，同步起始时间每天零点开始，目前采用 NGAS (Next-Generation Archive System, Andreas) 传输原始观测数据。用户可以登陆新疆台 Taurus 高性能计算系统，下载并处理数据，Taurus 与 Long-term Online Archive 间采用 56Gbps Infiniband 交换设备互连，用户处理后数据可根据需要进行归档、发布。GAVO（German Astrophysical Virtual Observatory）Servers 主要用于数据发布及处理后数据存储，原始观测数据元数据信息提取后，将被导入到相应的数据库中，为数据发布做准备。数据存储、Taurus 与 GAVO 服务器间采用 NFS 方式实现数据互操作。针对数据库数据，在新疆天文台本部及南山观测基地均有备份。目前新疆天文台奇台观测基地与台本部间已经建成 100Mbps MSTP（Multi-Service Transfer Platform 多业务传送平台）专线，为满足奇台前期建设及多种设备数据采集需要，已在奇台基地部署了一套 20TB 存储，这套存储同时也可以满足数据库备份的需要。

第三部分，由分别位于南山及奇台观测站的数据备份系统及数据发布平台组成。两套数据备份系统利用专线网络分别实现本部重要数据的远程、异地容

灾，数据发布系统以新疆天文台数据中心为基础实现观测数据基于虚拟天文台标准的发布、高效数据检索与数据获取[9]。

3、存储系统实现

3.1 存储技术介绍

存储系统根据服务器类型可分为封闭系统存储和开放系统存储，封闭系统主要应用于大型机，开放系统指基于 Windows[11]、UNIX[12]、Linux[13]等操作系统的服务器。开放系统存储又分为内置存储和外挂存储；外挂存储根据连接的方式分为直连式存储（Direct-Attached Storage，简称 DAS[14]）和网络化存储（Fabric-Attached Storage，简称 FAS[15]）；网络化存储根据传输协议又分为：网络接入存储（Network-Attached Storage，简称 NAS[16]）和存储区域网络（Storage Area Network，简称 SAN[17]），具体如图 2 所示。

- 封闭系统的存储
- 开放系统的存储
  - 内置存储
  - 外挂存储
    - **Direct-Attached Storage** 直连式存储**(DAS)**
    - **Fabric-Attached Storage** 网络存储**(FAS)**
      - **Network-Attached Storage** 网络接入存储**(NAS)**
      - **Storage Area Network** 存储区域网络**(SAN)**

图 2 **存储系统分类**

DAS 为当前最主要的应用模式，存储系统被直连到服务器，依赖服务器主机操作系统进行数据的 I/O 和存储维护管理，数据备份和恢复占用服务器主机 CPU[18]、系统 IO[19]等资源，数据流需要回流主机再到服务器存储，数据备份等操作约占用服务器主机资源的 20-30%，DAS 存储性能依赖于所接入的服务器设备。

---

[11] https://www.microsoft.com/zh-cn/
[12] http://www.unix.org/
[13] https://www.linux.org/
[14] https://en.wikipedia.org/wiki/Direct-attached_storage
[15] https://en.wikipedia.org/wiki/NetApp_filer
[16] https://en.wikipedia.org/wiki/Network-attached_storage
[17] https://en.wikipedia.org/wiki/Storage_area_network
[18] https://en.wikipedia.org/wiki/Central_processing_unit
[19] https://en.wikipedia.org/wiki/Input/output

NAS 存储也称网络附加存储，存储设备通过标准的网络拓扑结构添加到单台计算机或高性能计算系统。NAS 是文件级的存储方案，可以满足迅速增加存储容量的需求。支持即插即用、支持多计算平台，适用于 Unix/Windows 局域网，同时部署、应用非常灵活，但在备份过程中的带宽消耗较大。NAS 使用网络带宽进行备份和恢复，网络除了必须处理正常的最终用户数据传输外，还必须处理包括备份操作的存储磁盘 I/O 请求。

SAN 存储也称存储区域网络，通过光纤通道交换设备连接存储阵列和服务器主机，构建专用的存储网络，通过同一物理通道支持 SCSI[20]和 IP[21]协议，允许任何服务器连接到任何存储阵列，FCSAN[22]采用光纤接口具有更高的带宽，光纤接口支持超过 10KM 线路长度，使得物理上分离的、不在同一机房的备份存储变得容易实现。

基于对象的存储（Object-Based Storage，OBS[23] ），其核心是将数据通路（数据读、写）和控制通路（元数据）分离。基于对象存储（Object-based Storage Target，OST）构建系统，每个对象存储设备能够自动管理自身存储的数据分布，且具备一定智能。对象存储结构由对象、对象存储设备、元数据服务器、对象存储系统的客户端四部分组成。OBS 的网络带宽、IO 吞吐量、文件系统容量以及处理能力是随着存储节点的增加而同步线性增长，因而具有很好的性能和扩展性，存储节点可扩展、存储对象数可扩展性、存储对象空间也具有可扩展性。可以实现大规模的海量数据访问的高度并行，采用文件数据与元数据分离存储的机制，通过条带化技术将传统文件的数据分解存储到存储对象中；文件元数据则保存在元数据对象中，并具有一个全局唯一的对象标识以及一些文件属性信息。

存储局域网(SAN)和网络附加存储(NAS)是目前两种主流网络存储架构，而对象存储 OBS 是一种新的网络存储架构，OBS 综合了 NAS 和 SAN 的优点，同时具有 SAN 的高速直接访问和 NAS 的分布式数据共享等优势，提供了具有高性能、高可靠性、跨平台以及安全的数据共享存储体系结构。

[20] https://en.wikipedia.org/wiki/SCSI
[21] https://en.wikipedia.org/wiki/IP_address
[22] https://en.wikipedia.org/wiki/Fibre_Channel
[23] https://en.wikipedia.org/wiki/Object_storage

3.2 存储系统实现

综合考虑 DAS、NAS、SAN、OBS 技术的优缺点及目前新疆天文台观测数据的存储需要，最终采用基于对象的存储技术实现存储系统。系统以 Linux 下的 Lustre[24]为基础，Lustre 是基于对象存储的高性能分布式文件系统，源代码开放，使用基于对象的磁盘存储数据，元数据服务器为整个文件系统提供元数据服务。系统结构如图 2 所示，系统采用两套网络系统互连，56Gb Infiniband[25] 交换主要负责存储系统各服务器间链路，提供高速数据交换能力，千兆以太网实现管理。整个系统由两个元数据服务器（MDS）组成，两个 MDS 采用主备模式，数据实时同步，当主 MDS 故障时,备用 MDS 将接替工作，主备模式降低了系统故障率，保障了元数据信息正常访问。采用 3 台基于对象的存储设备（OST）作为目标存储节点，实现了 100TB 的可用存储空间。

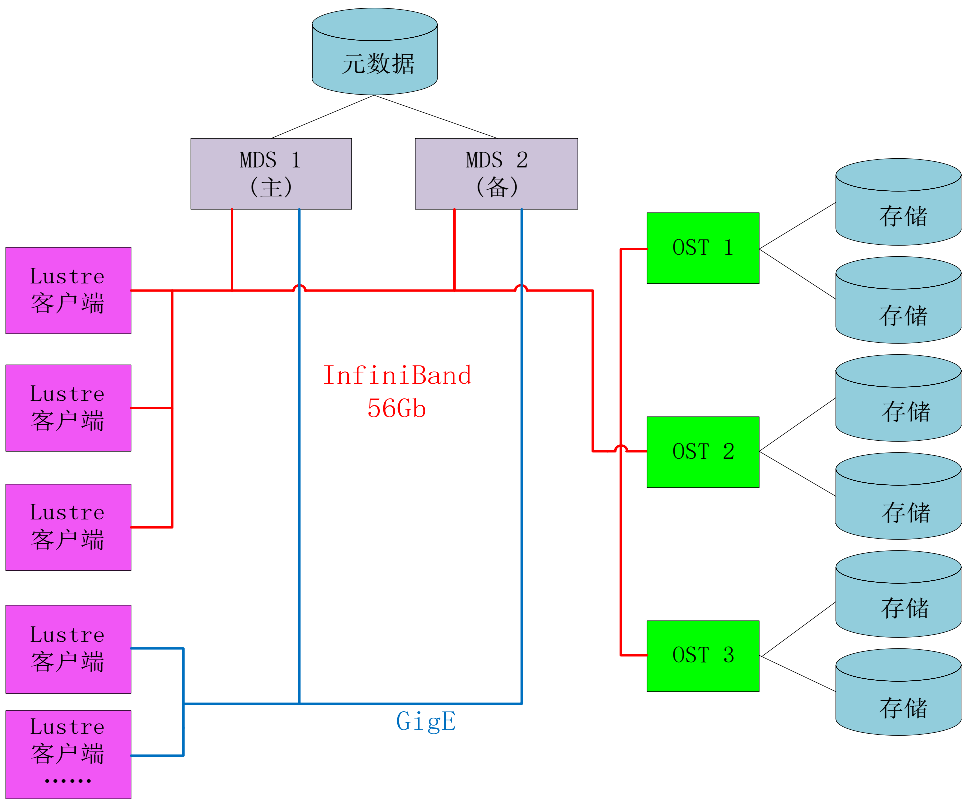

图 2 存储系统原理图

[24] http://www.lustre.org/
[25] https://en.wikipedia.org/wiki/InfiniBand

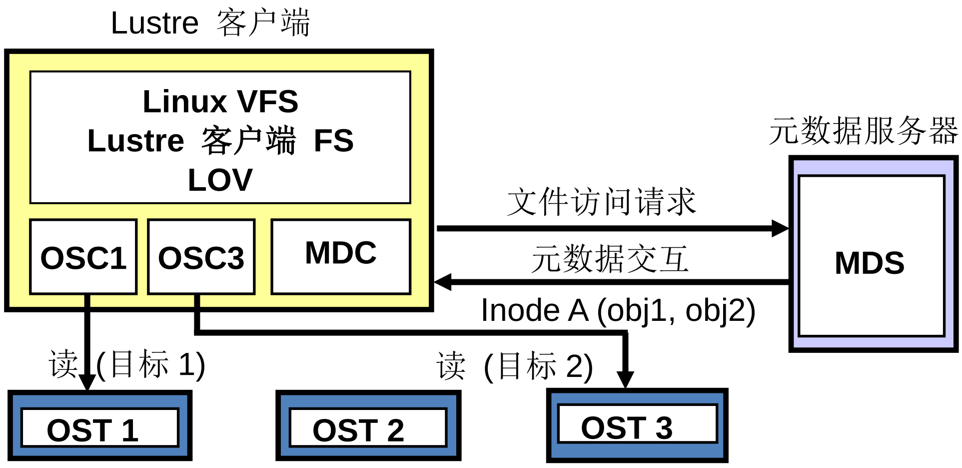

图 3 Linux 客户端并行读写示意图

所建设的集中式 Lustre 存储系统最终被连接到两台 I/O 服务器，I/O 服务器也采用主备模式，一台在线为客户提供服务，一台备份。对于 Linux 用户，需要安装相应的 Lustre 客户端软件，完成挂载后可以看到 100TB 存储空间。其数据访问示意如图 3 所示，Linux 虚拟文件系统[26]（VFS）通过同一套文件 I/O 系统实现 linux 中的任意文件操作，无需考虑其所在的具体文件系统格式，为能够支持各种实际文件系统，VFS 定义了所有文件系统均支持的基本的、概念上的接口和数据结构，Lustre FS(文件系统)提供 VFS 所期望的抽象接口和数据结构，将自身的文件、目录等概念在形式上与 VFS 的定义保持一致，实现两套系统间数据传递。逻辑存储卷（LOV）负责收集 OST 信息到单一卷中，用户的读写通过对象存储客户端（OSC）实现，OSC 得到用户的读写请求后，经过元数据客户端（MDC）查找元数据服务器（MDS）中对应的数据所在 OST 中位置并返回地址信息，OSC 得到 OST 的具体信息后实现并行数据读写。

3.3 存储性能测试

利用专业的存储性能测试工具 IOZONE[27]对所建设的系统读、写性能分别以单节点、多节点测试得到了相应结果。

1、单节点性能

测试命令：./iozone -a -g 64G -i 0 -i 1 -i 2 -f /home/iozone -Rb single.xls

[26] https://en.wikipedia.org/wiki/Virtual_file_system
[27] https://en.wikipedia.org/wiki/IOzone

参数说明：使用全自动模式，生成包括所有测试报告，使用的块大小从 4KB 到 16MB，最大测试文件 64GB, 测试节点来自文件/home/iozone，结果输出到文件 single.xls。

最终测试结果，分块大小为 8MB、文件大小为 8G，16G 左右取得最好性能，单点写入 420MB/S, 单点读 2.2GB/S。

2、多节点性能

测试命令：./iozone -a -g 64G -i 0 -i 1 -i 2 -f /home/nodes -Rb multi.xls

参数说明：使用全自动模式，生成包括所有测试报告，使用的块大小从 4KB 到 16MB，最大测试文件 64GB, 测试节点来自文件/home/nodes，nodes 文件中含有 8 个节点，结果输出到文件 multi.xls。

最终测试结果，分块大小为 8192KB、文件大小为 65MB 左右取得最好性能，多节点写入 960MB/S 左右，多节点读 5.1GB/S 左右，具体见图 4、图 5。

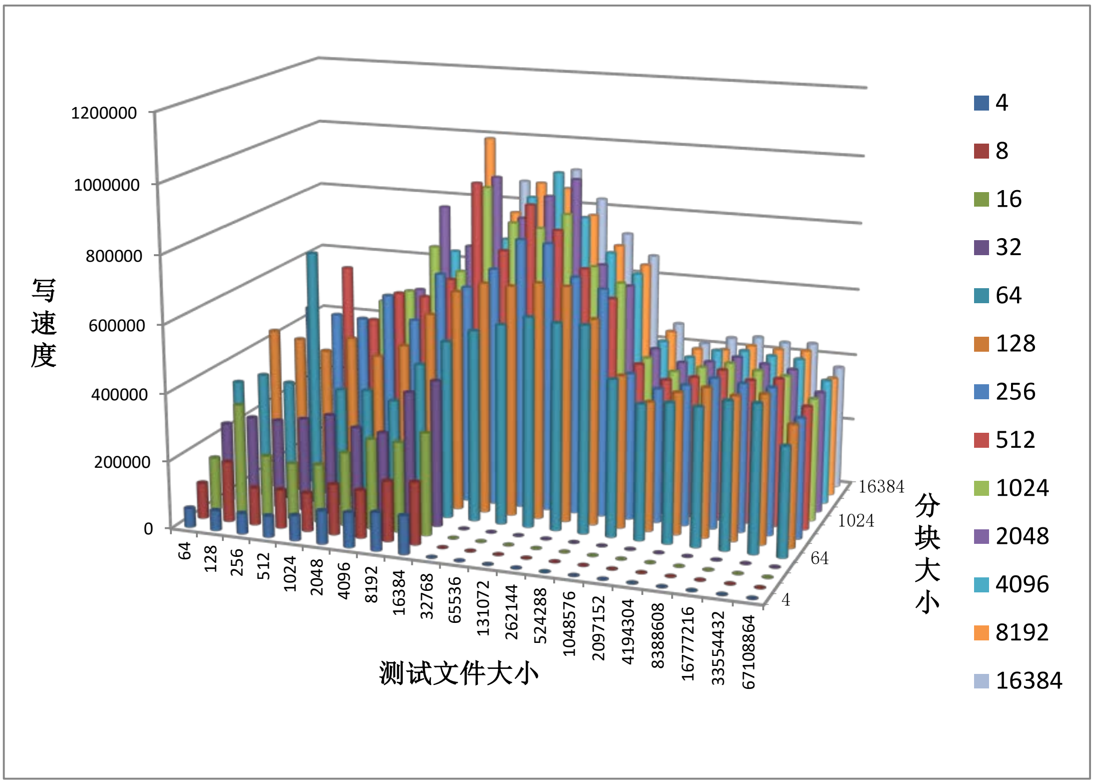

图 4 存储系统写速度（单位 KB）

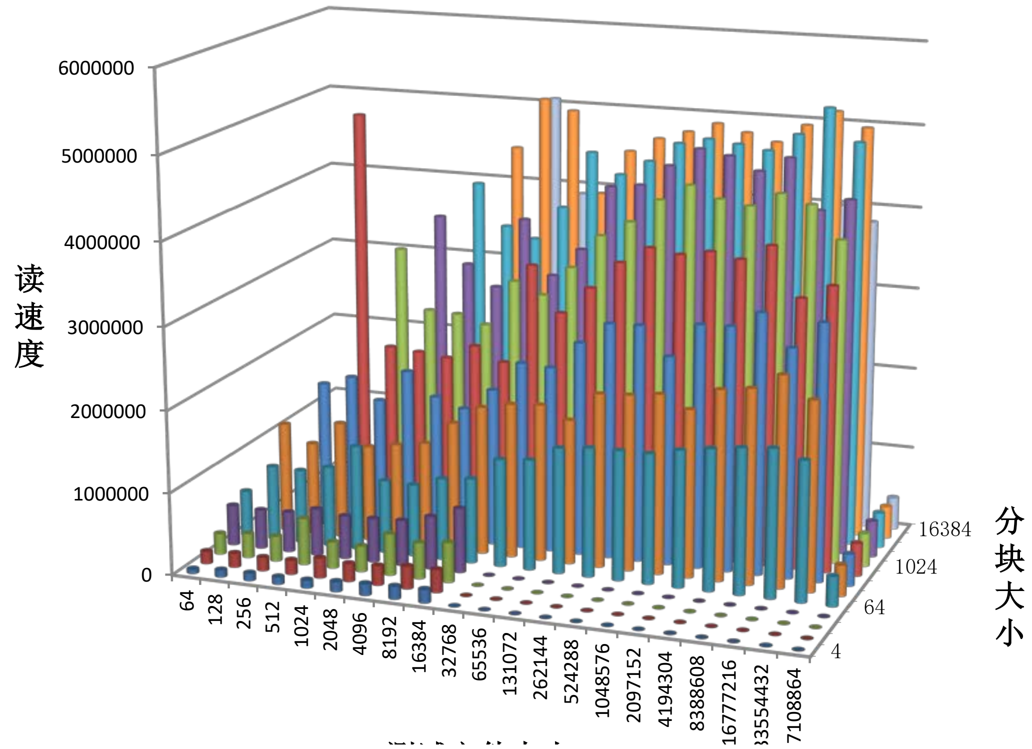

图 5 存储系统读速度（单位 KB）

4、结论

以新疆天文台的实际需求为基础设计并实现了 NSRT 观测数据的在线存储与备份系统，两套存储系统分别建设于新疆天文台本部与南山观测站，实现了远程、异地、容灾备份。对存储系统进行了读写性能测试，单节点、多节点读写速度目前可以满足 NSRT 数据管理需要。采用了基于对象的存储技术，所建设的存储系统具有良好的性能和可扩展性。

致谢:
NSRT 存储系统建设过程中的测试部分在新疆天文台数据中心及 Taurus 高性能计算系统上完成。

参考文献

[1] Tingay S J, Goeke R, Bowman J D, et al. The Murchison Widefield Array: the Square Kilometre Array Precursor at low radio frequencies[J].Publications of the Astronomical Society of Australia, 2013，30(30):109-121.

[2] 张海龙,王杰,王万琼, 等. 新疆天文台数据中心建设与应用[J]. 天文研究与技术,2017,14(1):94-102.

Zhang Hailong, Wang Jie, Wang Wanqiong, et al. Construction and application of the data center in Xinjiang Astronomical Observatory[J]. Astronomy Research and Technology, 2017,14(1): 94-102.

[3] Wang N, Manchester R N, Zhang J, et al. Pulsar timing at Urumqi Astronomical Observatory: observing system and results[J]. *Monthly Notices of the Royal Astronomical Society*, 2001,328(3):855–866.

[4] Yuan J P, Manchester R N, Wang N, et al. Pulse profiles and timing of PSR J1757−2421[J]. *Monthly Notices of the Royal Astronomical Society*, 2017,466(1):1234–1241.

[5] Hailong Zhang, Markus Demleitner, Na Wang, et al. Data retrieval from Xinjiang Astronomical Observatory’s Pulsar Data Archive [J]. Astronomy Research and Technology, 2016, 13 (4):473-480.

[6] Shirley Y L. The critical density and the effective excitation density of commonly observed molecular dense gas tracers[J]. *Publications of the Astronomical Society of the Pacific*, 2015,127(949):299–310.

[7] Liu B R, Liu X, Marchili N, et al. Two-year monitoring of intra-day variability of quasar 1156+295 at 4.8 GHz[J]. Astronomy & Astrophysics, 2013, 555(4):334-345.

[8] Liu X, Mi L G, Liu J, et al. Intra-day variability observations and the VLBI structure analysis of quasar S4 0917+624[J]. Astronomy & Astrophysics, 2015, 578: A34-A42.

[9] 张海龙,冶鑫晨,李慧娟,等. 天文数据检索与发布综述[J]. 天文研究与技

术,2017, 14(2):212-228.
Zhang Hailong, Ye Xincheng, Li Huijuan, et al. Astronomical data query and release review[J]. Astronomy Research and Technology, 2017,14(2): 212-228.

Xinjiang Astronomical Observatory NSRT Data Storage System

Zhang Hailong[1,2], Zhu Yan[1,3], Nie Jun[1,2], Yuan Jianping[1], Wu Gang[1], Liu Jun[1], Wang Jie[1], Wang Wanqiong[1], Ye xinchen[1], Tohtonur[1], Zhang Meng[1,3]

（1.Xinjiang Astronomical Observatory,Chinese Academy of Sciences,Urumqi 830011,China; 2.Key Laboratory of Radio Astronomy, Chinese Academy of Sciences,Nanjing 210008 China; 3.University of Chinese Academy of Sciences,Beijing 100049,China)

Abstract:

After years of observation, Xinjiang Astronomical Observatory(XAO) Nanshan 26 meters radio telescope (referred to as NSRT) had accumulated massive scientific data. A remote backup system was established for the online data storage of NSRT, this redundant storage system contains two storage clusters, one cluster was in XAO headquarters and another one located in Nanshan station, the real-time synchronization of NSRT data can be realized between two storage clusters. Based on the object storage technology, centralized Luster storage system was created for storage clusters, and the I/O performance test of luster systems was finished. Redundant storage system solved the online archive and data safety issue for NSRT data.

Keywords: Observational Data; Storage; Backup; VO; Data Safety