第 40 卷　第 1 期(251~259)　　中 国 地 震　　Vol. 40　No. 1
2024 年 3 月　　EARTHQUAKE RESEARCH IN CHINA　　Mar. 2024

吴峥,董翔,李杰飞,等. 2024. 地震数据分布式存储系统建设模式与服务效能研究. 中国地震,**40**(1):251~259.

地震数据分布式存储系统建设模式与服务效能研究

吴峥　董翔　李杰飞　曾薇　刘晓京

中国地震台网中心,北京　100045

摘要　我国地震监测系统历经了数字化、网络化和自动化的变革。随着地震监测站网规模的不断扩大,作为国家数据中心,中国地震台网中心面临数万地震台站观测数据处理和存储压力。本文基于前期国家地震烈度速报与预警工程、公共安全信息化工程(中国地震局)建设项目等建设经验,总结地震数据分布式存储系统的建设模式,并验证其合理性。同时以历史数据迁移归档业务为例,分析共享存储系统的服务效能,为海量观测数据的高效存储管理、共享数据存储系统建设提供参考。

关键词:　**地震数据　分布式存储系统　建设模式　服务效能**

[**文章编号**] 1001-4683(2024)01-0251-09　　[**中图分类号**] P315　　[**文献标识码**] A

0　引言

我国地震监测领域历经“中国数字地震观测系统的建设项目”、“数字地震观测网络项目”和“国家地震烈度速报与预警工程”(以下简称预警工程)等重大项目建设,形成了“国家中心-省中心-台站”的三级地震观测业务体系,实现了从模拟观测到数字化、网络化观测的巨大变革。各类观测仪器数量及观测数据量海量增长。以实时观测数据为例,预警工程在全国建设了 15899 个站点①,其中 1928 个基准站配备了 6 通道数采,3202 个基本站和 10769 个一般站配备了 3 通道数采,按照采样频率 100Hz、24 位计算,每通道每天的数据采集量为 100 次/秒×24bite/次×(24×60×60)秒/(8×1024)= 25.3125MB,经压缩处理后,每个 Miniseed 数据文件约 10MB,日产出文件约 1926×6+(3202+10769)×3 = 53469 个,故日增数据量约 10MB/个×53469 个/1024 = 522.16GB。前期建设的 1066 个测震站点,日产出文件 3000 余个,日增数据量约 30GB。地球物理场现有站点 972 个,在网运行 2966 套仪器②,按照秒采样、5 通道计算,日增数据文件 14830 个,数据量约 28GB。因此,在现有网络规模下,我国日新

[**收稿日期**] 2023-12-19　[**修定日期**] 2024-03-18

[**项目类别**] 公共安全信息化工程(中国地震局)建设项目、中国地震局 2023 年地震应急与信息学科重点任务(CEAITNS202302)共同资助

[**作者简介**] 吴峥,女,1991 年生,工程师,主要研究方向为地震信息基础设施设计建设。E-mail:wuzheng@seis.ac.cn

董翔,通讯作者,男,1983 年生,高级工程师,主要从事地震信息服务。E-mail:dongxiang@seis.ac.cn

① 中国地震局. 2018. 国家地震烈度速报与预警工程初步设计方案和投资概算报告.

② 国地震局监测预报司. 2023. 地震监测数据质量分析报告.

增数据文件为 53469+3000+14830=71299 个,日新增量为 522.16GB+30GB+28GB=580.16GB,年增量为 580.16GB×365 天≈210TB。上述数据整体呈现单个数据文件 MB 级、文件及数据增量庞大的特点。

中国地震台网中心(以下简称台网中心)作为国家中心,全局全量汇聚存储管理上述海量观测数据,目前已建成一套集中式存储(曾薇等,2011)和一套分布式存储系统,可用总容量约 7PB,承载了全国测震(含预警)及地球物理场自 2008 年以来的大部分原始观测数据和产品数据。同时,基于大数据等信息技术建设了地震数据管理系统,实现了地震观测数据实时或准实时汇聚存储管理,面向全国提供地震数据服务,并开展了一系列应用探索(陈通等,2022)。

1 分布式存储系统建设模式研究

1.1 分布式存储系统关键技术

分布式存储和集中式存储、蓝光存储、磁带库等均属于目前主流的数据共享存储方式。分布式存储是将大存储容量的机架式服务器作为存储介质,通过网络实现数据传输与交换,具有服务多元、统一管理、易于扩容等优势。目前市场上的分布式存储系统多是基于 Ceph 技术(陆华成,2023)实现的。Ceph 技术是一种开源分布式底层存储技术,具有良好的可扩展性,支持 PB 级数据存储,同时与 OpenStack、Kubernets 等技术有较好的兼容适配能力,在资源弹性扩容、数据自动容错、自动负载均衡等方面均有出色表现。Ceph 技术的去中心化架构有效避免了其他共享存储存在的管理节点单点故障问题。

分布式存储支持块、文件和对象存储等多种服务协议和方式。块存储(谭文贵等,2017)主要面向主机提供本地磁盘存储服务,支持 SCSI、iSCSI(吴怡之等,2015)等协议,还可以通过 RBD 协议直接对接虚拟机服务(洪亮,2017)。文件存储主要面向用户提供共享文件服务,具备完善的目录树结构,可读性强,用户可便捷地在不同主机访问同一存储空间,便于数据共享,目前较常用的协议有 NFS(蔡康宇,2022)、CIFS、SMB 和 FTP(郑韵等,2023)等。对象存储基于 S3 协议(李敏达,2022),面向应用提供存储接口服务,既方便共享又可快速访问,弥补块和文件存储的不足,适合于海量数据(杨力等,2023)和非结构化数据(田峰,2021)存储,同时可通过 S3 接口对接蓝光存储,实现从温热数据到冷数据的无缝衔接。

1.2 分布式存储系统架构设计

分布式存储系统主要依靠网络进行数据传输以及系统内部数据调度、同步等,网络性能直接决定了存储系统的服务性能,通过规划不同 IP 地址段进行网络隔离和识别,有效保证存储系统稳定可靠。服务器分为管理节点和存储节点,管理节点负责系统整体资源的协调统筹;存储节点需配置尽可能多的大容量数据存储盘存储数据,并按需配置一定数量的固态硬盘,用于缓存加速、元数据管理、索引数据管理等。通过在服务器上部署分布式存储软件,搭建分布式存储系统,实现块、文件、对象等多种存储资源池及存储服务,通过选择恰当的存储协议,对接上层地震业务应用,实现存储资源的最大化使用。分布式存储系统整体架构如图 1 所示。

1.3 分布式存储系统资源规划

为提供更好的服务体验,提高资源的使用效率,应考虑数据的分级分类存储,合理划分

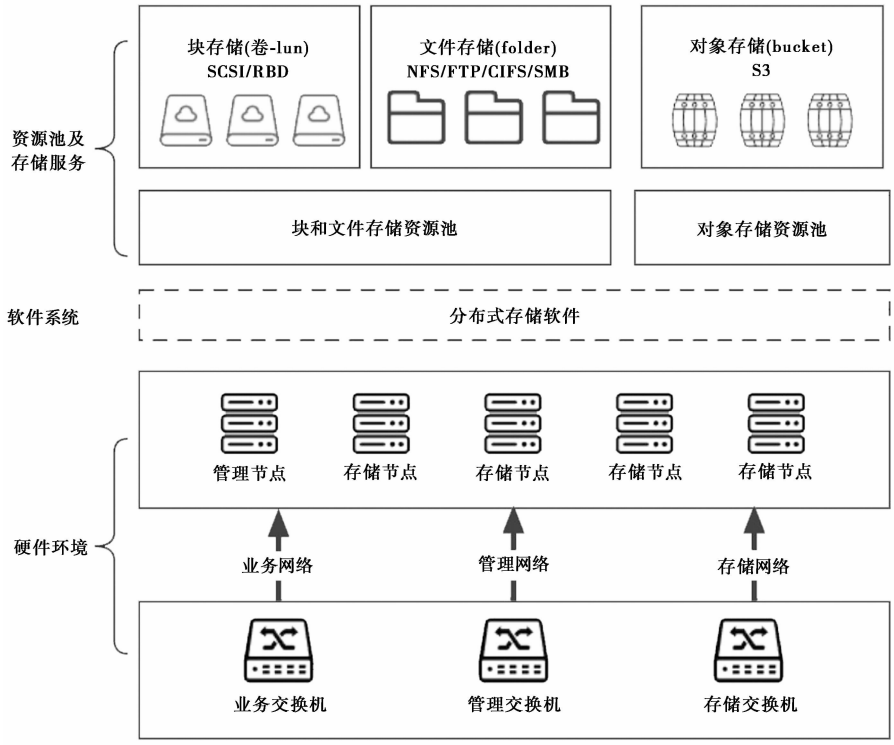

图 1　分布式存储系统架构

存储资源。热数据可选用机械硬盘或固态硬盘(如 SATA SSD 或 Nvme SSD),提升硬盘读写性能。温数据可选用机械硬盘,提升存储介质性价比,通过 S3 协议将其存储至蓝光存储介质中,以冷数据永久归档保存,降低机房能耗。同时,还需利用高性能固态硬盘搭建缓存加速池,用于存储元数据信息(块和文件存储)、索引信息和小文件缓存(对象存储),提高数据读写速率。

硬盘需在存储系统中做冗余配置,确保数据安全可靠。三副本方式在读写性能、可靠性、故障域范围等方面优于 EC 纠删码方式,但实际业务中还需综合考虑空间利用率、硬件成本等因素。缓存加速池和块存储对访问性能要求较高,采用三副本方式为宜;文件存储和对象存储对存储空间的需求更大,可采用 EC 纠删码方式。

台网中心须将实时汇聚的数据进行缓存、转发和在线快速处理,需要存储系统 7×24h“零”时延运转,一旦出现拥塞将会导致当前及往后一段时间的数据丢失,按照数据 7 天高速缓存、三副本存储计算,对于数据实时处理需求,所需热数据缓存空间约 580GB×7 天×3 副本=12TB;按照数据在线 3 个月计算,所需温数据存储空间约 580GB×92 天×3 副本=1.6PB。此外还须将原始数据以“台网. 台站. 仪器. 测项. 日历天”的文件方式进行存储管理,能够进行按需快速准确检索,该部分数据可通过 EC 纠删码方式分别存储到文件存储和对象存储中,满足不同业务应用,按照冗余 25%存储空间计算,年须存储量约210TB/0.75=280TB。

1.4 分布式存储系统网络连接

每台服务器配置 3 块双端口万兆光纤网卡。交叉选用不同网卡上的两个网口,采用 lacp 链路聚合模式绑定为 3 组 Bond(mode=4),并将网络划分为对外服务网络(网关网)、管理网络和内部网络(公共网和集群网),实现网络链路的冗余设计,扩大网络带宽,分担网络流量,避免单链路故障,提升系统稳定性。服务器上的端口使用情况、绑定情况和与交换机连接情况如图 2、表 1 所示。

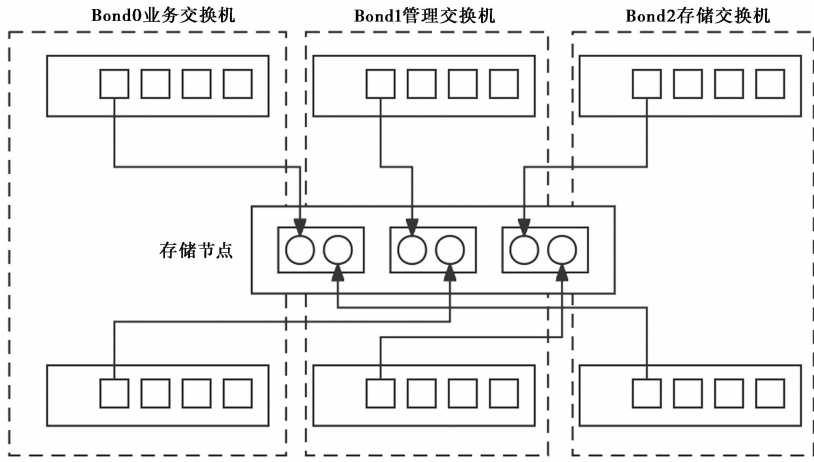

图 2 服务器网口连接交换机示意图

表 1 服务器端口使用表

网卡端口	Bond 名称	通信规则	接入网络	功能描述
网卡 1-1 口 网卡 2-1 口	Bond0 (配置子接口)	开放访问	网关网(2 个) 管理网	提供地震行业网和预警承载网的存储服务,系统运维
网卡 1-2 口 网卡 3-2 口	Bond1	封闭访问	公共网	存储系统内部数据调度
网卡 2-2 口 网卡 3-1 口	Bond2	封闭访问	集群网	系统内部磁盘监控及副本同步

备注:网卡 1-2 口,代表 1 号网卡上的第 2 个接口,以此类推。

1.5 分布式存储系统建设实践

台网中心分布式存储系统基于上述核心要点设计建设,共 25 台机架式服务器,每台机器配为 2 颗 intel 4214 CPU,192GB 内存,6 个万兆光口,2 块 480GB SATA SSD 配置 raid 1 模式,安装 Centos 7.8 操作系统,6 块 960GB SATA SSD 以 raid 直通模式做加速缓存,28 块 16TB SATA HDD 以 raid 直通模式做数据盘。其中 3 台服务器以三副本形式搭建块存储集群,各 11 台服务器以 EC8+2 的方式分别搭建文件和对象存储集群。同时,6 台万兆接入交换机通过两两堆叠实现数据链路冗余设计,通过 SDN 技术实现网关网、管理网、公共网和集群网的逻辑隔离。每台服务器上的 6 个万兆网口分别连接到一台接入交换机中,每台接入交换机再通过 40GE 端口上联至核心交换机,实现与局域网的互联互通。此外还有一套带外管理网络环境用于管理服务器硬件。详细的组网拓扑如图 3 所示。

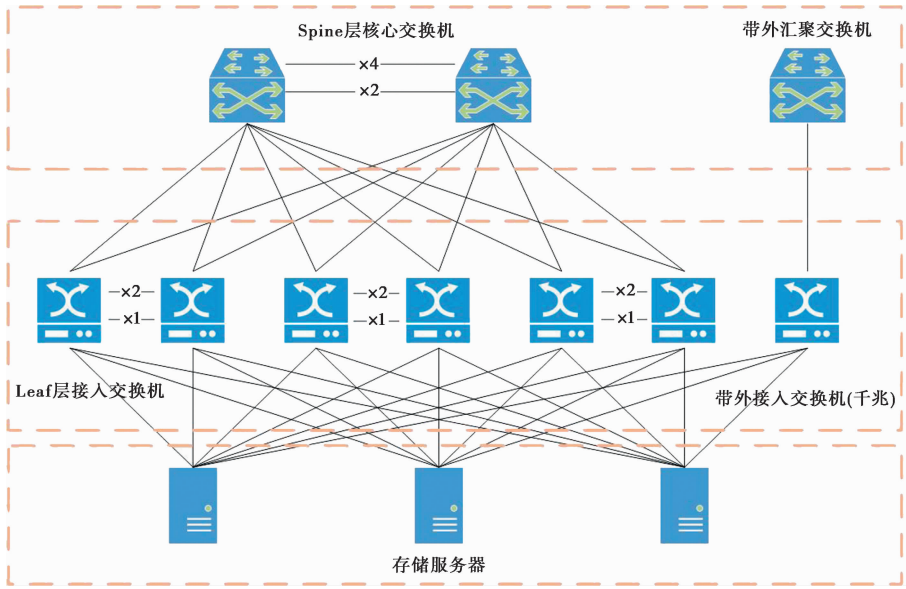

图 3　分布式存储系统组网示意图

2　分布式存储系统服务效能分析

2.1　实验目标

通过分析各存储方式在高并发环境下的吞吐、延时及带宽等性能指标，验证分布式存储系统设计的合理性。同时，利用真实地震数据，模拟历史数据迁移的业务场景，测试不同存储方式服务效能的差异。

2.2　实验环境及工具

实验所用的分布式存储为前文所述系统，集中式存储（NAS）采用磁盘阵列+8GB FC 光纤网络部署，分布式存储和集中式存储资源充足。

利用 1 台机架式服务器存放地震历史数据，该机器配为 2 颗 intel 6248R CPU，512GB 内存，2 块 480GB SATA SSD 做 raid 1，安装 Centos 7.8 操作系统，10 块 16TB SATA HDD 以 raid 5 模式做数据盘存放地震数据。该服务器通过万兆网络与两套共享存储系统相连。

压测客户端为 6 台与数据服务器配置相同的机架式服务器，安装 vdbench5.04.07 软件压测块存储和文件存储，安装 cosbench 软件压测对象存储。

2.3　实验结果

2.3.1　分布式存储系统压力测试

（1）块存储：利用 3 台压测客户端，每台设置 128 个进程，共并发 384 个进程。在块存储集群上创建块存储卷、客户端组、访问路径，并挂载到客户端上，选取每秒读写次数（iops）、网络带宽（BW）、网络延时（Lat）为测试指标，分别测试 4KB 小文件的随机读写性能和 1MB 大文件的顺序读写性能。测试结果见表 2。

通过存储平台自带的监控插件，查看各网络的带宽峰值数据，记录见表 3。

（2）文件存储：利用 5 台压测客户端，每台起 128 个进程，共并发 640 个进程。在文件

表 2 **块存储读写性能测试结果**

读写模型	iops	BW/(MB·s^{-1})	Lat/ms
4KB 随机写	38377.6	18.7	4.0
4KB 随机读	186493.6	91.0	1.6
1MB 顺序写	1495.2	186.9	150.0
1MB 顺序读	6586.32	823.3	55.0

表 3 **块存储带宽峰值记录**

网络平面	管理网/(MB·s^{-1})		网关网/(MB·s^{-1})		公共网/(MB·s^{-1})		集群网/(MB·s^{-1})	
	接收	发送	接收	发送	接收	发送	接收	发送
带宽峰值	0.4	0.5	889.2	1010.8	699.5	600.5	583.8	600.7

存储集群上创建 100TB 的 NFS 协议文件存储并挂载到客户端上,选取每秒读写次数(iops)、网络带宽(BW)、网络延时(Lat)为测试指标,分别测试 4KB 小文件的随机读写性能和 1MB 大文件的顺序读写性能。测试结果见表 4。

表 4 **文件存储读写性能测试结果**

读写模型	iops	BW /(MB·s^{-1})	Lat/ms
4KB 随机写	191382	93.4	2
4KB 随机读	203641	99.4	1
1MB 顺序写	8755	1094.4	10
1MB 顺序读	9026	1128.3	9

通过存储平台自带的监控插件,查看各网络的带宽峰值数据,记录见表 5。

表 5 **文件存储带宽峰值记录**

网络平面	管理网/(MB·s^{-1})		网关网/(MB·s^{-1})		公共网/(MB·s^{-1})		集群网/(MB·s^{-1})	
	接收	发送	接收	发送	接收	发送	接收	发送
带宽峰值	0.7	0.8	2250.3	1682.8	469.4	649.0	1112.1	1020.2

(3)对象存储:利用 6 台压测客户端,选取每秒钟能操作的对象数量(ops)、网络带宽(BW)、网络延时(Lat)为测试指标,分别测试 100 万 4KB 小文件和 100 万 4MB 大文件的上传、下载性能。测试结果见表 6。

表 6 **对象存储读写性能测试结果**

读写模型	ops	BW/(MB·s^{-1})	Lat/ms
100 万 4KB 小文件上传	13800	18.8	40
100 万 4KB 小文件下载	31100	22.5	20
10 万 4MB 大文件上传	1908	931.3	46
10 万 4MB 大文件下载	2284	1115.0	43

通过存储平台自带的监控插件,查看各网络的带宽峰值数据,记录见表 7。

表 7　　**对象存储带宽峰值记录**

网络平面	管理网/(MB·s^{-1})		网关网/(MB·s^{-1})		公共网/(MB·s^{-1})		集群网/(MB·s^{-1})	
	接收	发送	接收	发送	接收	发送	接收	发送
带宽峰值	1.0	1.1	1563.8	1290.4	812.5	759.1	767.2	795.6

2.3.2　历史数据迁移归档业务测试

选取两天历史地震数据，数据总量 1.1TB，共计 114500 个 Miniseed 文件，单个文件大小从 5MB 到 18MB 不等，平均每个文件约 9.6MB。将集中式存储(NAS)和分布式存储系统的文件存储分别挂载至数据服务器，通过复制命令将历史数据由服务器本地拷贝至共享存储。通过系统层 iftop 网络流量监控服务查看网络带宽情况，记录见表 8。

表 8　　**历史地震数据迁移网络带宽情况**

	集中式存储	分布式存储
平均网络带宽/(MB·s^{-1})	249.0	247.0

选取相同的数据，利用 java 编制基于 S3 对象存储协议的多线程历史数据迁移程序，实现数据从服务器本地迁移到共享存储中。通过 iftop 服务和存储系统监控软件查看相关性能指标，记录见表 9。

表 9　　**S3 对象存储历史地震数据迁移情况**

线程数	阻塞队列大小	平均网络带宽/(MB·s^{-1})	峰值网络带宽/(MB·s^{-1})	网络延迟峰值/ms	IO 大小峰值/MB
20	200	780.9	830.4	167.5	11.2
40	400	817.8	1044.5	279.6	12.4
50	500	719.1	825.1	247.5	11.8
60	600	616.4	706.5	420.9	27.1

当线程数为40 时，改变阻塞队列大小，记录实验数据见表 10。

表 10　　**阻塞队列 S3 对象存储历史地震数据迁移情况**

线程数	阻塞队列大小	平均网络带宽/(MB·s^{-1})	峰值网络带宽/(MB·s^{-1})	网络延迟峰值/ms	IO 大小峰值/MB
40	40	839.4	940.5	247.5	13.1
	100	854.5	1019.5	275.7	14.2
	200	820.3	951.6	251.4	13.2
	400	817.8	1044.5	279.6	12.4
	600	778.2	857.7	243.1	13.3

2.4 结果分析

2.4.1 分布式存储系统设计合理性

本次对块、文件、对象三种存储方式开展压力测试,结果表明:三种存储方式下,网络延时均在 150ms 以内,处于网络延时的合理范围内。网络延时越大,吞吐能力相对越弱,但吞吐能力与网络带宽间总体满足正相关关系,符合实际业务逻辑。

每个网络的理论网络带宽均为 20000Mbps(即 2500MB/s)。测试结果中,文件存储网关网的带宽峰值最大,为 2250.3MB/s,接近理论带宽上限的 90%。测试结果表明存储系统在不同存储方式、文件大小、并发数量等场景中,网络带宽均在理论值的合理范围内,未出现网络拥塞瓶颈。

网关网的带宽峰值普遍高于公共网和集群网,这是由于存储系统缓存池的加速作用,部分数据读写发生在缓存池,当缓存池写入容量超过 75%或当系统空闲时,数据才会通过公共网和集群网进行持久化存储,分散了公共网和集群网的网络流量。管理网流量基本维持在较低且平稳的水平,说明存储系统调度管理所需流量较小,且受业务变化影响小。

因此,分布式存储系统的设计较为科学合理。

2.4.2 历史数据迁移归档业务应用

本文将历史数据分别拷贝至集中式存储和分布式存储(NFS 和 S3 协议),模拟了地震历史数据迁移归档业务场景。结果表明,集中式存储和分布式存储系统的文件存储带宽使用基本相同,这是因为两者本质上都基于 NFS 协议。S3 对象存储的整体性能表现更优,这是由于 S3 对象存储简化了文件存储目录树、文件锁等功能,结构更加扁平,适宜存储无需修改的非结构化数据,适合于地震历史数据迁移归档业务。当线程数在 40 以下时,线程数与网络带宽呈正相关,线程数的增加有效激发了 CPU 处理能力;当线程数超过 40 时,网络带宽随线程数的增加而降低,这与数据服务器的 CPU 性能有关,过多线程导致上下文切换频繁,影响高并发执行速度。内存大小影响了阻塞队列大小的设定,阻塞队列为线程数的 2 倍左右为佳。由于本次实验仅使用了一台物理机,下一步还可采用不同类型 CPU、不同内存大小的数据服务器,来判定 CPU 和内存对迁移性能的影响。

3 结语

台网中心作为国家数据中心,面临着数万台站实时数据的汇聚、存储和管理需求,具有海量数据文件且数据密集增长的特点。分布式存储系统利用低成本的存储服务器搭建,具备高效、集约、一体化、弹性扩展等优势。本文通过实验证明了分布式存储系统的合理性,同时以数据归档业务为例,验证了 S3 对象存储技术能够在地震业务中发挥优势。地震监测、预警、预报等业务对共享存储的需求各有不同,在实际建设过程中需根据业务合理规划资源使用,充分发挥不同存储协议、存储介质的作用。

参考文献

蔡康宇. 2022. 基于 NFS 协议的计算机信息加密存储技术研究. 软件,**43**(1):148~150.

陈通,韩雪君,马延路. 2022. 时序数据库在海量地震波形数据分布式存储与处理中的应用初探. 中国地震,**38**(4):799~809.

洪亮. 2017. 开源分布式存储系统 Ceph 测试及在桌面虚拟化平台中的应用. 硕士学位论文. 广州:华南理工大学.

李敏达. 2022. 基于 Amazon S3 API 的分布式对象存储系统设计与实现. 硕士学位论文. 武汉：华中科技大学.
陆华成. 2023. 基于 Ceph 的混合存储性能优化研究. 硕士学位论文. 桂林：桂林电子科技大学.
谭文贵，黄英港，王琨. 2017. 一种基于 Ceph 提供弹性块存储的研究及实现. 信息通信，(10)：216～217.
田峰. 2021. 基于 HDFS 的海量小文件存储系统的研究与实现. 硕士学位论文. 西安：西安电子科技大学.
吴怡之，田双杰，周宇艳. 2015. 基于 iSCSI 的软件定义存储局域网研究. 计算机科学，42(增刊Ⅰ)：253～255，259.
杨力，陈建廷，向阳. 2023. 基于 HBase 的工业时序大数据分布式存储性能优化策略. 计算机应用，**43**(3)：759～766.
曾薇，杨乐，谭颖. 2011. 网络存储技术在地震数据存储中的应用. 震灾防御技术，**6**(3)：335～342.
郑韵，王青平，郑超，等. 2023. 基于 FTP 协议的地震应急产品共享系统的设计与实现. 华南地震，**43**(2)：77～82.

Construction Mode and Service Efficiency of Seismic Data Shared Storage System

Wu Zheng, Dong Xiang, Li Jiefei, Zeng Wei, Liu Xiaojing
China Earthquake Networks Center, Beijing 100045, China

Abstract Earthquake monitoring system of China has experienced the innovation of digitalization, cyberization and automation. With the expansion of earthquake monitoring station and network, China Earthquake Networks Center (CENC), as the national earthquake data center, is facing serious pressure of processing and storing data from tens of thousands of earthquake monitoring stations. This paper mainly focuses on the key points in constructing the distributed storage system based on experiences from Earthquake Early Warning Project and Public Security Information Project (CEA subproject), and trying to demonstrate its advantage and rationality. This paper also analyzes the service capacity of shared server system with the case example of history data transfer and archive business. The results of this paper will provide further reference for the storage and management of mass earthquake data efficiently and the construction of shared data storage systems in different departments.

Keywords: **Earthquake data**; **Distributed storage system**; **Construction mode**; **Service efficiency**