文章编号：1001-1749(2024)02-0235-07

面向大规模集群作业并发规模的数据库连接池优化技术

师 伟，王向辉，林茂春，侯红军，程 实
（中国石油东方地球物理公司，涿州 072751）

摘 要：数据库作为地震勘探处理软件系统的多学科数据存储管理的核心组件，在底层支撑着处理作业的读写访问需求。在当前，随着地震勘探处理技术的飞速发展，大规模集群下处理作业的并发规模也快速扩张，底层数据库采用常规的读写一体化部署方式难以支撑大规模并发作业的读写请求。笔者针对地震勘探大规模集群数据资料处理作业并发场景，提出一种“1＋N“读写分离部署方式的数据库连接池优化技术，设计了基于数据库服务器节点信息的资源分配器，对并发作业的数据库读写请求进行了合理的优化，并在实验室环境和实际地震勘探数据资料处理生产中进行了验证，能够支撑大规模集群下处理作业长事物、高并发等特征的数据库访问请求。

关键词：数据库；读写请求；集群作业；并发规模
中图分类号：P631.4 **文献标志码**：A **DOI**：10.3969/j.issn.1001-1749.2024.02.13

0 引言

随着地震勘探采集技术的不断深入应用，人工地震勘探单位面积采集数据量快速上涨，部分三维地震勘探项目数据量已经超过 PB 级。数据量的快速增长给地震资料处理软件带来了全方位的冲击，计算单元（CPU/GPU）、存储系统、网络通信、数据库等多种企业级高性能核心组件逐步应用到了大型工业化地震资料处理中。在计算单元部署方面，通常采用集群（Cluster）系统来对高性能计算节点进行管理和维护。集群是通过网络资源实现独立的计算节点相互连接的计算机集合[1]，能够有效地组织、管理计算资源，提高处理地震作业效率。如中石化某物探院，其部署的浪潮集群，整体管理约 300 台高性能计算节点来进行地震资料处理[2]。东方物探研究院，其采用开源软件 Zabbix 进行集群节点数据的可视化分布式监控，整体管理超过 2 000 台计算节点来进行地震资料处理[3]。因此，采用大规模高性能计算集群来应对地震勘探计算密集型的业务需求已经成为整体趋势。特别是在地震资料处理面向 PB 级大型数据密集型场景时，采用千节点规模的集群和 GeoEast 并行批量作业引擎软硬件相结合的方式来作为生产解决方案[4]。方案中大规模批量作业被分配到集群各个计算节点来执行，每个作业的模块通过网络访问数据库实现读写访问。多个模块在同一时间内，通过封装的接口读写访问数据库资源，极其容易造成数据竞争，从而引发数据库过载，严重限制了批量处理地震作业的并发规模。

在国内开源数据库应用研究中，在大规模集群能源数据库优化使用中，李俊等[4-5]提出采用 pgpool 对数据库瞬时访问进行负载，能够提高部分并发规模。而国内地震勘探数据处理在解决大规模集群作业对数据库服务器所带来的访问压力，通常采用增加物理内存和调参优化的方式进行优化。地震数据处理作业是对 PB 级地震数据资料进行计算操作，在作业启动和作业结束时会对数据库进行工区、网格等关键数据读写。对数据库的访问方式和互联

收稿日期：2022-09-13
基金项目：中国石油集团科学研究与技术开发项目（2021ZG03）
第一作者：师伟（1989—），男，本科，主要研究方向为地震勘探数据应用研究，E-mail：shiwei_01@cnpc.com.cn。

网在线数据访问存在较明显区别。因此笔者提出一种针对大规模集群批量处理地震作业并发场景，数据库微服务集群采用“1＋N”(1 个写数据库服务器和 N 个读数据库服务器的组合)读写分离部署方式，计算节点采用连接池动态管理当前节点的数据库客户端连接资源，通过对数据库的访问连接、SQL 脚本合理优化，实现了数据库对批量作业并发规模的支撑。最后在地震资料处理实验室环境和实际生产环境下对优化方案的可行性进行了验证。

1 数据库并发

目前，主流的地震勘探处理软件如 Omega2014、GeoEastV3.X 等采用的是 C/S 架构。以 GeoEast 为例，其客户端通常都是高性能计算节点[5-6]，主要用来执行处理地震作业对地震资料进行偏移计算等；其数据库服务端主要负责存储部分核心数据，主要有速度 TV 对、地震数据部分索引、工区、测线、网格等。其中处理地震作业是运行一个处理任务的进程，其任务内容整体包含输入、输出模块(geodiskin、geodiskout)，振幅均衡模块(AmpEqu)等 GeoEast 专属模块[7]。在客户端节点上模块通过软件平台业务逻辑层对数据库进行读写网络访问。举例来说，一块 1T 的叠前地震数据，预计采用 20 个计算节点来进行处理，每个节点理论上执行 10 个任务，每个任务相当于一个进程，该作业共计 200 个进程同时执行。在实际执行过程中，由于集群节点的闲忙状态有差异，任务平均分配不能够最大限度地用集群的计算力，而是会根据节点的处理能力来动态分配，如图 1 所示。因此各节点对数据库服务器的请求也是有差异的，整体主要表现在以下三个特点：

1)长连接

地震作业的模块对数据库的访问是通过 SOCKET 实现的，通过建立 SOCKET 连接与数据库服务器保持通讯。由于部分处理模块的特殊性，需要与数据库保持及时的数据交互，所以部分接口采用长连接来实现，在作业开始后，各模块开始依次顺序运行，每个模块调用业务逻辑层接口与数据库服务器建立 SOCKET 连接，在作业模块运行的全生命周期都是保持连接状态的，而部分作业执行时间从几分钟到几天。随着作业的进程快速批量执行，模块持有的连接仍然保持，数据库服务器资源会渐渐不足。

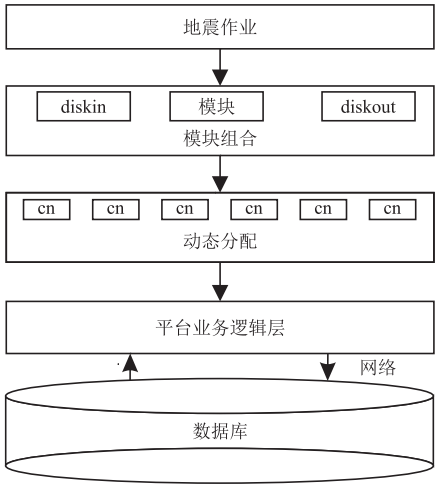

图 1 作业访问数据库示意图

Fig. 1 Seismic jobs operation database

2)多连接

一个大的地震数据作业会被并行作业引擎分割成多个任务，地震数据会根据索引和用户操作，被分为若干个独立的数据块。这些任务提交给集群各节点去执行，每个节点执行完当前任务，向控制节点请求新任务继续执行。这个过程中，作业对每一个输入文件进行访问时候，每一个文件都会占用一个连接，如图 2 所示，作业输入文件体量越大，数量越多，连接数占用也就越多。而数据库服务器能够承受的最大连接数是有限的，这些连接在模块运行生命周期里不会消亡。

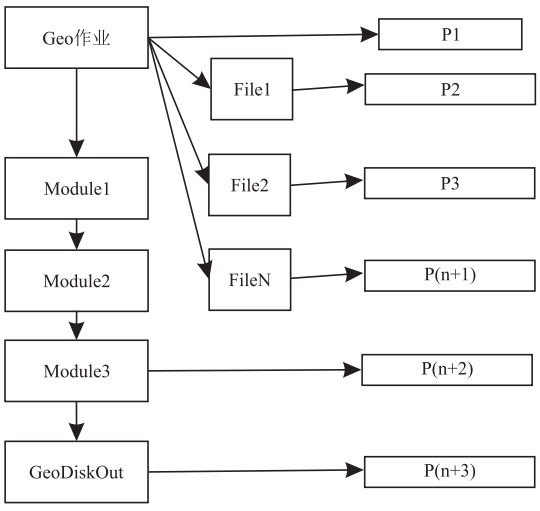

图 2 作业访问数据库连接示意图

Fig. 2 Seismic job's module operation database

3)峰值高负载

作业模块在集群节点通过数据库客户端访问数

据库服务器，比如 PostgreSQL（PG）客户端 libpq，在作业批量发送执行的一个时间点，模块对服务器的访问瞬间暴增，多个客户端实例通过 SOCKET 与服务器建立连接，完成认证，进行地震数据项目、工区、索引等信息的访问。通常在这个时间段，客户端的访问具有峰值高负载，抢资源的特点。客户端通过 SOCKET 连接主要对数据库服务器进行数据访问，如图 3，某时刻负载直接拉满，获取相关信息后开始执行其他模块，作业后续模块连接状况在某个阶段上下均衡波动。

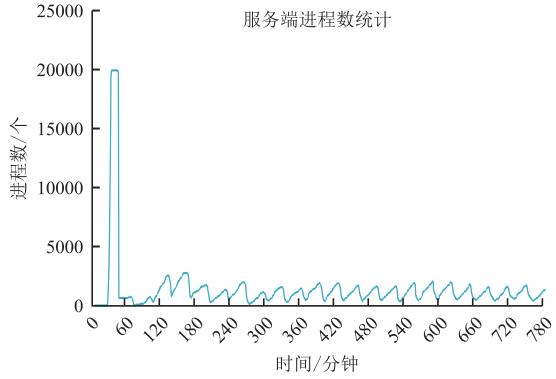

图 3　某时段数据库服务器连接进程数统计分钟图

Fig. 3　Minute chart of database server connection processes in a period

以上三个特点，让工业化地震处理作业的效率无法实现最优化，处理业务峰值无法最大化。根据中石油东方物探研究院某中心工业生产反馈，随着“两宽一高”等采集技术的深入，处理地震资料数据的体量不断增大，对数据库的读写访问需求不断增加。在地震数据常规偏移处理过程中，大规模的偏移作业会造成作业错误和系统死机，其中硬件故障、软件崩溃等因素引起的系统死机只占到很小一部分，但是处理作业瞬时启动，大量作业同时访问数据库服务器，造成对数据库服务器的连接资源抢占，使软件的作业阻塞或者排队等候，进而形成死机现象。数据库服务器的连接资源限制了集群偏移作业的并发规模。数据库的连接数等资源支撑并发的能力成为当前限制地震作业并发规模的主要瓶颈之一。地震作业在大规模集群下对数据库并发访问不同于常规的互联网数据库并发。处理地震作业的客户端基本都是高性能节点，对数据库服务器的访问具有一定的时段性，客户端在作业启动和结束的瞬间频繁和数据库服务器进行读写访问交互，后续过程中对数据库服务器仅读访问，而互联网客户端如手机端、平板端等民用设备，性能普通，通常采用对数据库服务器优化方案即可。所以大规模集群作业并发规模的优化既要考虑数据库服务器优化又要考虑客户端的调整。而在业内，特别是大型工业化软件的稳定性和业务峰值规模是要对数据库服务器进行测试和优化的。比如 JMeter、httperf、S-Clinet 等 Web 端常用负载测试工具，通常是对数据库服务器进行负载测试，缺乏关注高性能客户端负载的调整和优化[8]。

2　设计方案与实践

当前，地震勘探处理软件的数据库服务器采用的是读写一体化部署方式。在一个高性能服务器上部署一套数据库，同时支撑着多个生产项目的运行。而生产项目所使用的集群的计算节点数量可以根据项目的要求和进度动态地调配扩展[9]。随着地震数据单个体量的不断增大，单进程的数据处理耗时通常需要数十小时。处理整个工区数据，所需要的时间往往要超过项目预计的最优时间。因此，在地震资料工业化提速提效处理时，采用大规模集群并行执行地震作业已经是客观事实。而数据库服务器的连接负载、瞬时吞吐量和批量地震作业的并发规模之间存在着一定差距。笔者针对大规模集群地震作业并发场景设计了服务器和客户端联合的优化方案，服务器采用“1＋N”的读写分离部署方式，客户端采用动态管理数据库客户端连接资源的连接池技术，通过对接口的读写分离的定制对并发作业的数据库读写请求进行了合理的优化。

2.1　设计方案

目前，实现读写分离的方式有很多，主流的方式大体两个思路。一是基于用户自定义应用的自主代码逻辑设计，将读操作和写操作分发到各自的读、写数据库服务器上去执行。但是必须要处理好读写数据库的数据同步问题，业务代码要自己手动实现，主要优化在客户端的业务逻辑层以及底层的代码。二是采用开源的读写分离中间件来实现，完全不需要考虑客户端的优化，在服务端部署分发中间件，释放出访问接口，读写操作由中间件来实现。但是这种方式的优化，对服务器资源不能够精准把握，过程中分发逻辑和业务急需不能够定制满足。因此，本方案整体采用的是思路一，在部分环节加入地震勘探处理用户的定制需求。

基于大规模集群环境下作业并发规模的数据库

连接池优化整体方案如图 4 所示，数据库服务器采用微集群结构，数据库采用开源 PG 数据库进行部署。客户端采用 PG 客户端 libpq 与数据库服务器建立连接和读写访问。通过对 libpq 的封装，实现数据库访问驱动(PGdriver)的定制。通过配置文件(Config)实现对服务端读写分离集群的定向访问。基于客户端的动态连接池(Cpool)，让模块(module)能够快速高效地获取和释放数据库连接资源。

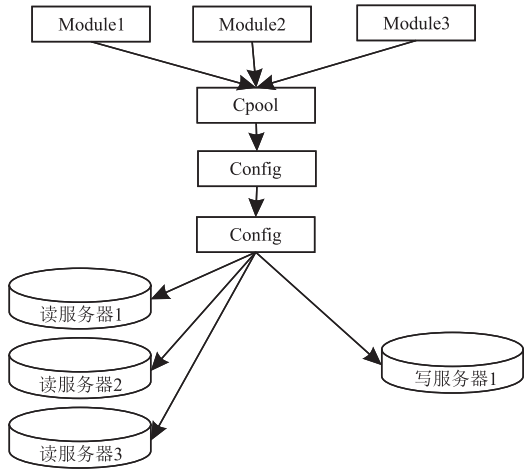

图 4 连接池整体优化方案示意图

Fig. 4 Schematic diagram of overall optimization scheme of connection pool

在服务端的优化，是基于地震作业生产业务的访问需求来设计的。根据生产数据库的日志分析来看，地震作业对数据库的访问请求中，读请求远远大于写请求，特别是在批量作业执行启动的时间段，几乎全部都是读取请求。在作业运行过程中，作业模块对数据库读请求是写请求的数十倍。因此，在服务器采用 1 个写数据库和 N 个读数据库，通过流复制实现写入信息的同步。其中，写数据库仅负责作业对工区、网格、观测系统数据的写入，不承担任何作业的读数据请求。读数据库主要负责作业对地震勘探项目、工区等观测系统信息和地震数据索引的查询服务。

在客户端首先对作业输入文件进行了优化。对于地震作业对于文件和数据库连接的绑定关系进行了解绑，如图 5 所示，对于 GeoEast 地震作业的输入文件，通过作业模块获取数据库连接，仅获取一个，用后释放。从业务逻辑底层进行优化，减少数据库资源占用，不论批量作业引擎执行对大体量地震数据分割为多少个数据文件，通过一个连接实现地震数据存储路径，观测系统等信息的获取。

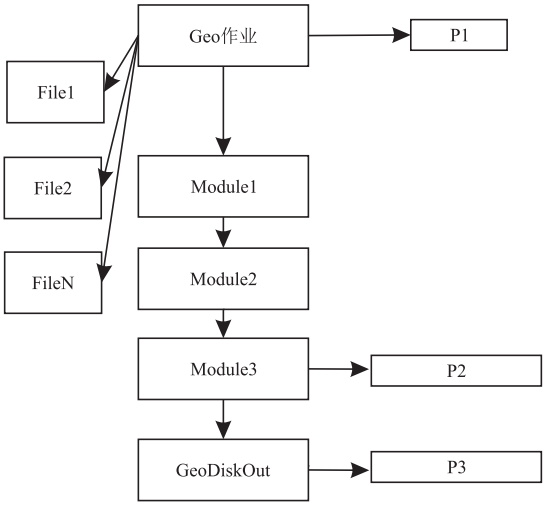

图 5 客户端优化方案示意图

Fig. 5 Client optimization scheme

针对地震勘探处理软件中数据库服务器的吞吐量和软件开发许可协议等因素，对外部开源以及多种客户端(libpq)的中级间进行调研，优化方案整体采用轻量级连接池技术和基于开源 PG 数据库客户端 libpq 的自研连接技术。

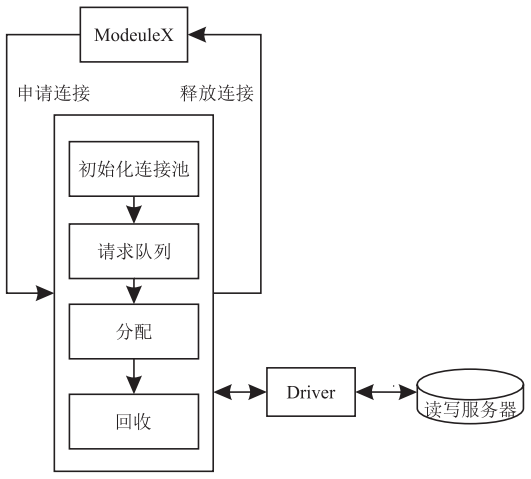

图 6 连接池整体优化方案示意图

Fig. 6 Schematic diagram of overall optimization scheme of connection pool

动态连接池 Cpool 是采用二维链表来实现的，对每个客户端初始化一个 Cpool 实例，释放固定的数据库连接资源，以二维链表的形式存储。通过释放连接池接口，让客户端对连接进行访问。访问请求进入请求队列，根据现有连接数动态分配。在优化方案中，应用模块不允许长期持有数据库连接这一宝贵资源，即使应用有长期持有的需求，也会通过连接池自动连接和释放，当其它应用模块发起数据

访问请求时，应用接口也不需要重新建立连接，而是从连接池中获取已有连接，减少接口的访问时间（数据连接建立耗时较长），提高接口的响应效率。能够支持多进程或者多线程的数据访问，以实现连接资源的最大程度的共享。

2.2　实验验证

实验集群在 IBM 集群进行优化方案的测试。测试环境有 256 个计算节点，每个节点配置 2 路 8 核 CPU，128 GB 内存，文件存储采用磁盘阵列。数据库服务器节点配置一致，采用读写一体的数据节点和 2 个读 1 个写的读写分离的数据服务器集群做对比组。实验数据采用某地震处理项目数据体，测试数据共 16 TB。

测试地震作业最基本要满足数据的读写效率和正确性，并且，能够有效提升批量作业的并发规模。对处理业务中常用的静校正和去噪作业进行正确性测试，如图 7，作业任务包含数据输入、输出模块，用户自定义模块等。采用后台日志监控，对模块对数据库的连接资源调用情况进行实时监控。通过不断递增输入文件的个数，对比优化前后作业对数据库连接数实时的占用数量，如图 8 所示。针对地震数据大数据、多输入文件的作业，不管输入文件个数多少，连接数始终只占用 1 个，相对于优化前的方案，大幅度减少了资源的占用，优化后的方案对于作业过程中对于数据库连接的访问情况趋于稳定，作业任务的每个模块能够正常的执行，并能够正常有效地反馈处理成果。从计算节点客户端优化了数据库连接的使用效率。

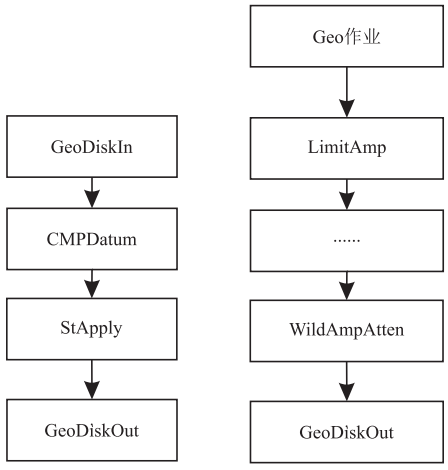

图 7　静校正和去噪作业

Fig. 7　Static correction and noise elimination

针对地震数据大数据、多输入文件的作业的特点，对优化方案前后进行了批量作业时间消耗对比测试。在保持输入文件总数整体基本不变的情况下，通过不断递增作业的文件输入个数，对执行地震作业去处理完所有文件的运行时间进行测试对比，如图 9。

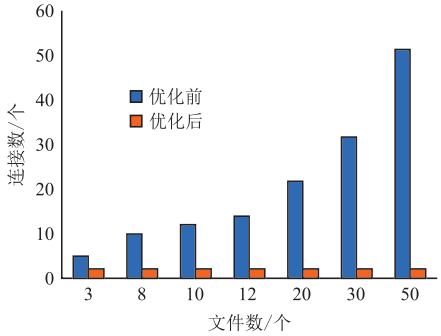

图 8　优化前后作业输入文件连接数对比图

Fig. 8　Comparison of the number of connections between the input files before and after optimization

大规模偏移作业需要计算节点和数据服务器节点配合执行，作业执行消耗的时间体现的是集群计算力。在地震数据处理生产过程中，发送包含多个小文件的单独作业和包含一个文件的批量作业都是基本的应用场景。测试模拟实际应用场景，在测试集群对 451 个作业，每个作业输入 1 个文件和 1 个作业输入 451 个文件进行对比测试，优化前后作业执行完所消耗的时间基本一致，能够让集群的计算力充分发挥。并且在不考虑网络延迟、磁盘 I/O 读写速度的情况下，仅从数据库连接资源的优化单方面看，大规模偏移作业不再出现死机的情况。

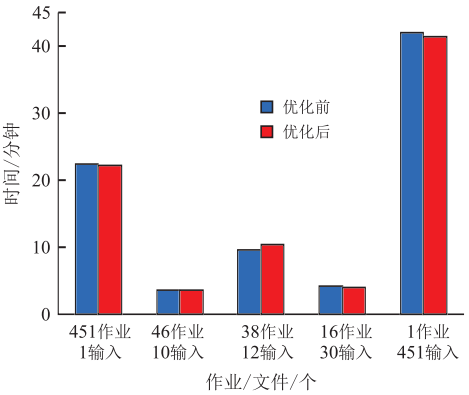

图 9　优化前后同类型作业时间监测对比图

Fig. 9　Comparison of monitoring of the same type of operation time before and after optimization

采用读写分离，化长为短，变静为动的服务端和客户端联合优化方案，让地震作业的模块对数据库连接资源的使用率大大提高，在相对规模的集群下进行批量作业的运行测试，其客户端数据库连接资源使用率大大提高，其作业整体运行效率基本保持不变。特别是针对大规模集群批量并行作业执行引擎在对大体量、多文件地震数据的并发规模生产应用中，更需要有实际的生产测试。

在东方物探研究院某中心，采用约 2 000 节点的计算集群和“1＋N”的读写分离部署数据库服务器进行相应的数据测试。每个计算节点去执行地震作业，作业至少包含输入、输出模块。预计每个节点执行 10 个作业，2 000 节点共计执行约 20 000 个作业，批量作业引擎根据节点实际运行状态动态分布作业任务，作业模块通过客户端的连接池实例与数据库服务器进行访问通讯。同体量级别的作业规模在优化前的生产应用中是无法实现的。在作业执行的峰值时间段内，是对服务器数据库和客户端连接池等优化方案的性能检验。在作业任务执行过程中，对其中某时间段的读写数据库进程数据进行统计如图 10，能够直观地看到，大规模的作业峰值并发能够达到 20 000 的数量级。其中读库的性能可以稳定支撑峰值的模块数据访问，大大分担了写库的查询负载。写库进程维持在某个区间内平衡震荡，整体支撑的并发规模得到有效地扩展。

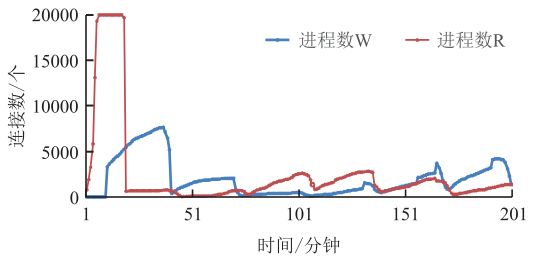

图 10 “1＋N”读写分离部署服务端进程支撑时段图

Fig. 10 “1＋N” Read write separation deployment server process support period

3 结论

地震勘探数据处理与企业级数据库的性能优化紧密联系。随着物探采集技术的发展，业务峰值的并发规模可能会越来越大，对数据库的负载均衡是一个新的挑战。笔者提出了面向高性能客户端和常规数据库服务器的联合优化方案，通过连接池和读写分离部署实现了地震作业并发规模的提升，根据作业模块的请求和客户端的性能，保证了作业运行的正确性和效率，有利于充分利用大规模集群的资源。

参考文献：

[1] 李军茹，侯红军，林茂春，等. 两宽一高海量地震数据的存储与处理方案[J]. 石油工业计算机应用，2018，26(2)：6-10
LI J R，HOU H J，LIN M C，et al. Storage and processing scheme of two wide and one high massive seismic data [J] Computer Application in Petroleum Industry，2018，26 (2)：6-10. (In Chinese)

[2] 姜游，陈军，黄骏. 高性能计算机在地震资料处理中的应用[J]. 计算机工程与科学，2009，31(S1)：328-330.
JIANG Y，CHEN J，HUANG J. Application of high-performance computer in seismic data processing [J]. Computer Engineering and Science，2009，31 (S1)：328-330. (In Chinese)

[3] 陈涛，梁妍，曹士炳，等. 基于处理解释集群的整体CPU利用率监控方法[J]. 信息与电脑(理论版)，2021，33(20)：13-15.
CHEN T，LIANG Y，CAO S B，et al. Monitoring method of overall CPU utilization based on processing interpretation cluster [J]. Information and Computer (Theoretical Edition)，2021，33 (20)：13-15. (In Chinese)

[4] 文佳敏，赵长海，侯红军，等. GeoEast海量地震数据高效处理技术[J]. 石油工业计算机应用，2016，24(3)：12-19.
WEN J M，ZHAO C H，HOU H J，et al. GeoEast massive seismic data efficient processing technology [J] Computer Application in Petroleum Industry，2016，24 (3)：12-19. (In Chinese)

[5] 李俊. 基于PostgreSQL集群的能源数据采集存储系统的研究与实现[D]. 广州：华南理工大学，2013.
LI J. Research and implementation of energy data acquisition and storage system based on PostgreSQL cluster [D]. Guangzhou：South China University of Technology，2013. (In Chinese)

[6] 黄代军. 地震资料处理软件Omega2014的配置及运行[J]. 江汉石油职工大学学报，2017，30(4)：16-18，37.
HUANG D J. Configuration and operation of seismic data processing software Omega2014 [J]. Journal of Jianghan Petroleum Staff University，2017，30 (4)：16-18，37. (In Chinese)

[7] 杜吉国，孙孝萍，陈继红，等. Lustre并行文件系统在地震数据处理中的应用[J]. 物探装备，2013，23(5)：294-299.

DU J G, SUN X P, CHEN J H, et al. The application of lustre parallel file system in seismic data processing [J]. Geophysical Equipment, 2013, 23 (5): 294-299. (In Chinese)

[8] 晋文明，颜硕彦，钱巨. 大规模负载生成客户端影响因素研究[J]. 计算机与现代化，2020(8)：76-81.
JIN W M, YAN S Y, QIAN J, et al. Research on the factors affecting the large scale load generation client [J]. Computer and Modernization, 2020 (8): 76-81. (In Chinese)

[9] 尹龙，张卫华，程实，等. 大规模计算机集群在地震勘探资料处理中的应用探讨[J]. 计算机时代，2016(8)：1-3.
YIN L, ZHANG W H, CHENG S, et al. Discussion on the application of large-scale computer cluster in seismic exploration data processing [J]. Computer Age, 2016 (8): 1-3. (In Chinese)

Application of database pool technology for job concurrent scale in large-scale cluster environment

SHI Wei, WANG Xianghui, LIN Maochun, HOU Hongjun, CHENG Shi

(Bureau of Geophysical Prospecting INC., China National Petroleum Corporation(BGP), Zhuozhou 072751, China)

Abstract: As the core component of multidisciplinary data storage and management of seismic exploration processing software system, the database supports the reading and writing access requirements of processing jobs at the bottom. With the rapid development of seismic exploration processing technology, the concurrent scale of processing jobs under large-scale clusters is also expanding rapidly. The e's conventional read-write integrated deployment mode makes it challenging to support the read-write requests of large-scale concurrent jobs. This paper presents a concurrent scenario for data processing of large-scale cluster seismic exploration, The database adopts "1+N" The connection pool optimization technology of "read-write separation deployment mode" designs a resource allocator based on the node information of the database server, reasonably optimizes the database read-write requests of concurrent jobs, and has been verified in the laboratory environment and actual seismic exploration data processing production. It can support processing long things and high concurrency under large-scale cluster Database access requests.

Keywords: database; read write request; cluster job; concurrent scale