DOI:10.20033/j.1003-7241.(2025)08-0111-04

基于改进哈希算法的PostgreSQL数据库容灾备份研究

裴旭斌，陈泽堃，陈旺平

（国网浙江省电力有限公司信息通信分公司，浙江 杭州 310016）

摘　要：介绍哈希算法的基本概念和操作流程，提出一种改进算法。将输入数据分块处理，并为每个块分配一个哈希计算单元，实现并行计算以缩短计算时间。性能测试包括存储速率、数据备份量和故障节点影响等方面的评估。实验结果显示，改进后的哈希算法在处理大量数据时具有较高的存储效率和备份效率，但在大规模数据备份方面效率稍显不足。此外，随着故障节点数量增加，算法的运行时间也相应增加。

关键词：改进哈希算法；并行计算；数据备份；SIMD指令

中图分类号：TP309.3　　**文献标识码：**A　　**文章编号：**1003-7241(2025)08-0111-04

Research on PostgreSQL Database Disaster Recovery Backup Based on Improved Hash Algorithm

PEI Xubin, CHEN Zekun, CHEN Wangping

(State Grid Zhejiang Information and Telecommunication Branch, Hangzhou 310016, China)

Abstract: This paper introduces the basic concept and operation flow of hash algorithm, and then proposes an improved algorithm, which divides the input data into blocks and assigns a hash computing unit to each block, to realize parallel computation and shorten the computation time. Performance tests includes evaluations of storage rates, data backup volumes, and the impact of failed nodes. The experimental results show that the improved hash algorithm has higher storage efficiency and backup efficiency when dealing with a large amount of data, but it is slightly less efficient in large-scale data backup. In addition, the running time of the algorithm increases with the increase of the number of faulty nodes.

Keywords: improved hash algorithm; parallel computing; data backup; SIMD instructions

0　引言

容灾备份是大规模数据库不容忽视的重要问题，此方面学者对于数据安全与备份解决方案进行了广泛的研究。张脉琪研究了基于云平台的企业远程办公文件分布式存储方法[1]。秦慧娟对资源属性进行处理，并利用B+树建立索引库，并设计相应的检索算法[2]。胡呈梅采用B/S架构设计整体框架体系，结合区块链技术设计登录验证模块、档案信息处理模块和查询模块，以增强数据库访问控制能力。通过哈希算法+RSE算法对信息进行加密处理，并设置三重安全访问机制来强化访问者行为审查力度[3]。黄一英等提出了通过并行的沃尔什—哈达玛转换和新哈希算法提高音频检索的完整性和鲁棒性[4]。罗莉霞等提出了一种基于MapReduce的并行相似矩阵构造算法，用于大规模数据集[5]。杨粟等提出了一种无监督并行哈希图像检索模型，使用CNN学习图像特征并使用相关度距离构建伪标签矩阵[6]。叶昱等探讨了医院门诊系统对应急策略的信息化储存、检索与匹配系统应用，并分析其备份和更新系统[7]。花强等提出了一种基于随机森林的哈希检索算法，实验证明该方法在MNIST和CIFAR-10数据集上有效[8]。熊舒羽等提出了一种基于深度哈希算法的图像并行检索方法，使用CNN建立图像特征和哈希码提取模型[9]。韩丽容等研究了海洋观测图像的深度哈希检索及FPGA在轨实现，采用CNN提取特征并训练哈希算法[10]。这些文献展示了在不同领域的研究中普遍存在的趋势和特点，特别是数据存储方面的研究呈上升趋势，反映了对于提高系统可伸缩性和可用性的兴趣。基于上述现有研究成果，本文以电力通信网络为例，探讨如何利用改进的哈希算法进行PostgreSQL数据库的容灾备份。

1　哈希算法及其改进

1.1　哈希算法综述

目录哈希树是一种用于高效组织和查询大量数据的数据结构。电力通信网络中，目录哈希树节点由七个元组构成，包括父节点指针、数据信息标志位、哈希列表文件夹、遍历次数、数据文件或文件夹路径、文件或文件夹

*基金项目：浙江省计划项目(2022KY957)

收稿日期：2024-05-24

名称以及子文件夹名称。具体表示为

<*father,hash,hashMap,flag,path,
fileName,childSET> （1）

加密的哈希算法应用于数据库时，使用连续的 r 个块，每个块包含 $b=k\cdot w$ 个比特，每个块又包含 k 个连续的字，每个字有 w 个比特。而布隆过滤器总共有 m 个比特，因此有：

$$m=r\cdot b=r\cdot k\cdot w \quad (2)$$

在数据库中应用加密哈希算法时，数据被分成连续的块，每个块包含一定数量的比特，进一步分为连续的字，每个字包含若干比特。布隆过滤器总共包含一定数量的比特，插入过程中，元素信息被编码并存储在随机选择的块中。哈希函数用 h_1 于选择块，并在选定块内选择特定 k 个比特进行设置。每个哈希函数 h_0 与一 $word[i]$ 关联，并且只能在对应字内操作。

1.2 容灾备份数据操作

网络节点部署需要考虑数据层、列表层和资源层三个水平，其中数据层位于网络最顶端，负责集中存储通信数据；列表层位于网络中部，负责接收来自数据层的数据并与其他资源建立联系；资源层位于网络最底端，负责调整待处理的信息参量并二次调试相关通信节点所处位置。

基于此，考虑数据备份中上传与下载的操作流程中，备份操作涉及选择随机块 block[h0(e)]存储元素 e 信息，并设置其中某些 k 个比特为 1。查询时使用相同方法 h_0 判断元素是否存在于集合中。

1.3 算法改进

为提高哈希计算速度，UFBF 使用 SIMD 指令执行哈希函数。SIMD（single instruction，multiple data）指令允许并行执行多个相同的操作。改进方法包括初始化种子、加载到 SIMD 寄存器、执行并行哈希计算函数以及将结果写入内存。这个算法并不生成新的哈希函数，而只是使用 SIMD 指令对传统哈希函数进行并行计算的改进。这种方式通过并行计算提高了哈希函数的效率，使得哈希算法在处理大量数据时更加高效。

具体的改进涉及将传统哈希函数中的操作转化为 SIMD 操作。

其中，

val 代表哈希值的中间计算结果。

OP 表示通用的逻辑算术运算。

vr_a 代表将 a 复制为 p 份的 SIMD 寄存器。

v_OP 代表执行逻辑算术运算 OP 对应的 SIMD 操作。

通过这种方式，改进哈希算法可以将大量的数据处理任务分布到不同的网络节点上，降低边缘节点对集中存储节点的影响，提高数据传输完整性水平，有效地适应并保持高效的容灾备份能力。

1.4 其他性能优化措施

动态调整块大小可以使各处理单元的计算负载大致相等，避免某些处理单元过早完成计算而其他处理单元仍在繁忙。在并行计算中，有效处理数据之间的依赖关系是必要的。利用硬件加速器如 GPU 或 FPGA 执行哈希计算可以显著提升计算速度。内存访问优化，并且线程池管理一组预先创建的线程可有效提高系统整体性能和响应速度。处理海量数据时，采用分布式计算框架可以显著提高整体计算速度和系统吞吐量，并提供良好的横向扩展性和容错性。

2 性能测试

本文进行了两项主要的性能测试：储存速率测试和数据备份量测试，并检验了不同故障节点数量条件下的平均运行时间。这些测试旨在评估本文的算法在不同条件下的性能表现，并提供有关算法效率的深入理解。

2.1 储存速率测试

本文生成了一系列不同大小的数据集，进行了储存速率测试，评估算法在处理不同数据输出量时的储存速率。记录了每个数据集的储存时间，并使用此信息计算了储存速率。

对较小规模样本的测试结果如图 1 所示。

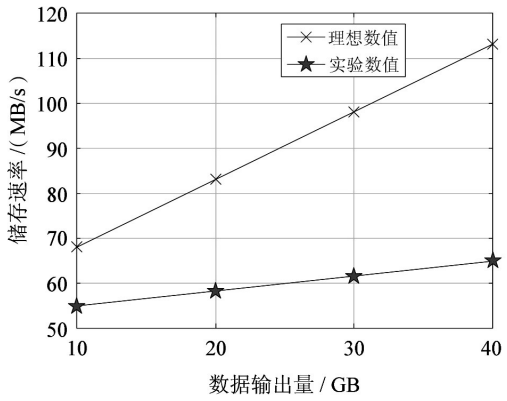


图 1 小规模样本的储存速率测试

如图 1 所示，小样本规模下理想数值和实验数值均呈现稳定增长态势。

扩大样本规模进行测试如图 2 所示。

实验结果显示，随着数据输出量的增加，储存速率呈现出一种近似线性的增长趋势。当数据输出量为 10 GB 时，储存速率为 68 MB/s；而当数据输出量为 1 000 GB 时，储存速率达到了 1 553 MB/s。这种趋势表明，本文的算法在处理大量数据时的储存效率较高。

本文的算法采用了并行哈希计算策略，使得处理多个数据块时的效率得到了提升。另外，本文也注意到实验数值与理想数值之间存在一定的差距，这可能是因为

在实际的硬件环境中，存在一些影响性能的因素，例如内存带宽限制、CPU缓存命中率等。

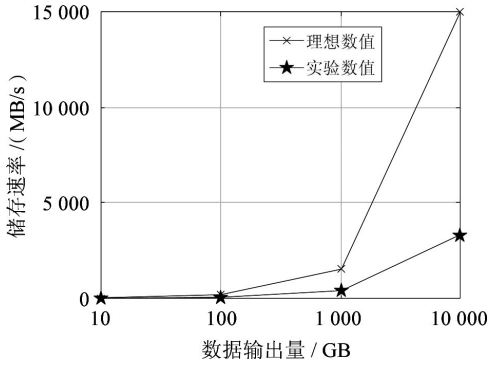


图2 大规模样本的储存速率测试

2.2 数据备份量测试

本文创建了一系列不同大小的数据备份集，进行了数据备份量测试，评估算法在处理不同数据备份量时的备份时间，并使用此信息计算了备份速率。

对较小规模样本的测试结果如图3所示。

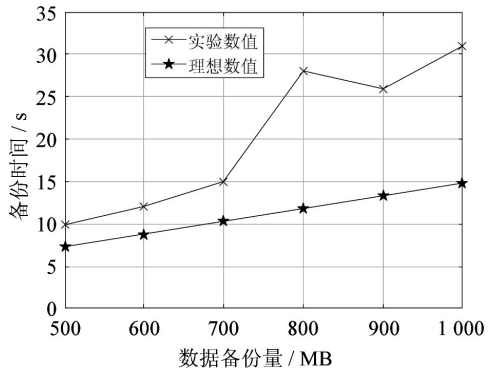


图3 小规模样本的数据备份量测试

如图3所示，随着数据备份量增长，则其相应备份时间增长，但这一增长有一定波动性，且整体备份耗时相对较长。

扩大样本规模进行测试如图4所示。

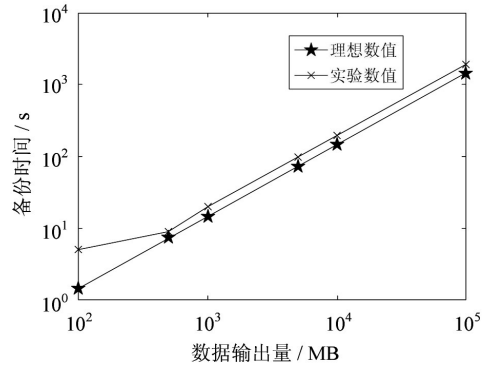


图4 大规模样本的数据备份量测试

实验结果显示，随着数据备份量的增加，备份时间呈现出一种近似线性的增长趋势，这一趋势的规律性增加。当数据备份量为100 MB时，备份速率为1.47 s；而当数据备份量为1 000 MB时，备份速率为14.71 s。这种趋势表明，本文的算法在处理大量数据备份时的运行效率相对较低。

本文的算法在处理大量数据备份时，由于内存访问模式、CPU缓存命中率等因素的影响，导致性能下降。另外，实验数值与理想数值之间存在的差距可能是因为在实际的硬件环境中，存在一些影响性能的因素，例如内存带宽限制、CPU缓存命中率等。

总的来说，通过这两项性能测试，发现本文的算法在处理大量数据时的储存效率和备份效率都比较高。然而，在处理大量数据备份时，由于受到硬件环境等因素的影响，备份效率相对较低。针对这些问题，计划在未来进行更多的优化工作，以提高算法在处理大量数据备份时的性能表现。

2.3 故障节点影响

在不同故障节点数条件下的测量结果如图5所示。

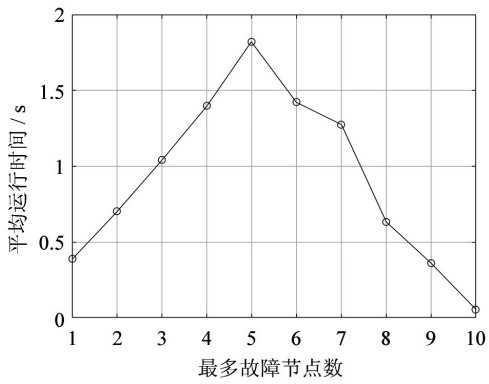


图5 故障节点影响

当网络中的故障节点数增加时，本文的算法需要花费更多的时间来处理数据并恢复网络的完整性。这是因为随着故障节点数的增加，需要进行的备份和恢复操作也会相应增加，从而增加了算法的运行时间。

可能由于本文的算法在处理故障节点时采用了较为保守的策略，例如在检测到故障节点时需要进行额外的数据备份和验证操作。另外，实验数值与理想数值之间存在的差距可能是因为在实际的硬件环境中，存在一些影响性能的因素，例如网络延迟、带宽限制等。

总的来说，这个实验结果表明本文的算法在网络出现故障时，能够有效地处理数据并恢复网络的完整性。然而，随着故障节点数的增加，算法的运行时间也会相应增加。因此，计划在未来优化算法的性能，例如通过采用更高效的故障处理策略或优化数据备份和恢复的过程，以减少算法在处理故障节点时的运行时间。

3 结束语

本文提出了一种改进的哈希算法，该算法通过提高

哈希计算的并行性来提高整体性能。性能测试基于储存速率、数据备份量和故障节点影响测试，评估了算法在不同条件下的性能表现。结果显示，改进后的哈希算法在处理大规模数据时表现出较高的储存效率和备份效率，随着故障节点数的增加，算法的运行时间也会相应增加，但仍能应对故障环境维持备份工作的有效进行。这些发现为改进哈希算法在大规模数据处理中的应用提供了参考，展示了其潜力和局限性。

参考文献：

[1] 张脉琪. 基于云平台的企业远程办公文件分布式存储方法[J]. 自动化技术与应用, 2024, 43(3): 112-115.

[2] 秦慧娟. 基于B+树的多媒体教育资源层次化索引模型[J]. 自动化技术与应用, 2023, 42(3): 101-104, 128.

[3] 胡呈梅. 基于区块链技术的铁路项目档案管理信息系统设计[J]. 自动化技术与应用, 2023, 42(2): 114-116,125.

[4] 黄一英, 邓开发, 邬春学. 基于空间自适应哈希算法的并行双音频指纹技术[J]. 计算机应用研究, 2022, 39(7): 2076-2080, 2086.

[5] 罗莉霞, 蒋盛益. 基于MapReduce的相似矩阵并行构造[J]. 计算机工程与设计, 2021, 42(5): 1368-1375.

[6] 杨粟, 欧阳智, 杜逆索. 基于相关度距离的无监督并行哈希图像检索[J]. 计算机应用, 2021, 41(7): 1902-1907.

[7] 叶昱, 邓豪, 刘漫. 门诊应急策略信息化匹配平台自动更新设计[J]. 微型电脑应用, 2021, 37(9): 100-102.

[8] 花强, 郭欣欣, 张峰, 等. 基于随机森林的哈希检索算法[J].计算机科学与探索, 2019,13(7): 1174-1183.

[9] 熊舒羽, 毛雷, 刘畅. 基于深度哈希的批量图像并行检索方法[J]. 重庆理工大学学报(自然科学), 2018, 32(1): 188-194.

[10] 韩丽容, 李鹏, 王廷伟, 等. 海洋观测图像的深度哈希检索及FPGA在轨实现[J]. 海洋科学, 2018, 42(1): 58-64.

作者简介：裴旭斌(1973—)，男，本科，高级工程师，研究方向：数据库、DEVOPS等。

(上接第101页)

3 结束语

本文以铁路通信继电保护测试异常诊断作为研究重点，在大量文献资料的支持下，提出基于深度学习的铁路通信继电保护测试异常诊断技术研究。在深度学习技术的支持下，实现了对继电保护测试异常的高效诊断。并在异常诊断技术开发完成后，针对铁路通信的实际情况设计大量实验内容，结果证明了提出的异常诊断技术的可靠性和可行性。虽然本文在铁路通信继电保护测试异常诊断上取得了不错的研究成果，但是现有研究水平仍有一些不足之处。因此，在以后的研究发展上，从智能化、安全化入手，展开更深层次的研究与讨论，不断完善继电保护测试异常诊断技术。

参考文献：

[1] 丁晓兵, 陈朝晖, 周兆庆, 等. 继电保护装置单粒子效应的测试方法与失效率研究[J]. 电力系统保护与控制, 2022, 50(24): 167-171.

[2] 刘晶儒, 胡志云, 叶景峰, 等. 喷气发动机湍流燃烧场激光定量诊断技术研究进展[J]. 推进技术, 2022, 43(3): 9-26.

[3] 王桥. 地表异常遥感探测与即时诊断方法研究框架[J]. 测绘学报, 2022, 51(7): 1141-1152.

[4] 彭成, 李凤娟, 蒋金元. 基于深度学习的故障诊断与预测方法综述[J]. 现代电子技术, 2022, 45(3): 111-120.

[5] 彭桂喜, 袁思遥, 孙昊, 等. 继电保护装置检验测试智能体的研究与实现[J]. 电力系统保护与控制, 2022, 50(3): 174-179.

[6] 王相, 杨耀忠, 何岩峰, 等. 基于深度学习的油井工况智能诊断技术研究及应用[J]. 油气地质与采收率, 2022, 29(1): 181-189.

[7] 江晓春. 数据融合技术在数字化变电站继电保护防误中的应用[J]. 通信电源技术, 2021, 38(23): 167-169.

[8] 宋晓川, 郑宽昀, 李俊臣, 等. 基于计量自动化系统的电能计量异常诊断技术研究[J]. 电网与清洁能源, 2022, 38(4): 110-116.

[9] 李玉敦, 耿玉杰, 管荑, 等. 基于边缘计算的二次设备动态诊断技术的研究与应用[J]. 电子器件, 2021, 44(5): 1176-1182.

[10] 朱保军, 咸日常, 范慧芳, 等. WRSR与改进朴素贝叶斯融合的变压器故障诊断技术研究[J]. 电力系统保护与控制, 2021, 49(20): 120-128.

[11] 何云良, 裘愉涛, 吴路明, 等. 基于5G通信的继电保护技术研究[J]. 电力系统保护与控制, 2021, 49(7): 31-38.

[12] 郭强亮, 王开锋, 李辉, 等. 重载铁路MCX多媒体调度通信技术应用研究——以靖神铁路为例[J]. 铁道标准设计, 2022, 66(6): 159-162, 173.

[13] 王立宪, 马宏忠, 戴锋. GIL机械故障诊断与预警技术研究[J]. 电机与控制应用, 2021, 48(8): 106-113.

[14] 赵武智, 樊国盛, 邬小坤, 等. 基于IEC 61850的智能变电站二次回路安全隔离研究[J]. 自动化技术与应用, 2024, 43(6): 141-145.

[15] 朱少斌, 许素安, 马宗彪, 等. 基于BSO-BPNN模型的电能计量装置异常诊断方法研究[J]. 中国测试, 2022, 48(1): 141-146.

[16] 朱鹏宇, 蔡新忠, 徐士元, 等. 面向电力通信网边缘计算的缺陷诊断研究[J]. 电子技术应用, 2021, 47(4): 30-35, 45.

作者简介：张志刚(1970—)，男，硕士，高级工程师，研究方向：铁道牵引供电。