文章编号:1671-4598(2019)03-0103-04 DOI:10.16526/j.cnki.11—4762/tp.2019.03.020 中图分类号:TP301 文献标识码:A

存储器中易失性用户大数据并行处理控制系统

梁 霄[1,2]
(1.青海警官职业学院,西宁 青海 810000;2.河海大学,常州 江苏 213000)

摘要:传统的并行处理控制系统在处理存储器中易失性用户大数据时,对CPU的利用率很低,导致处理控制工作精密度差;为了解决此问题,设计了一种新的大数据并行处理控制系统,分别对系统的硬件和软件进行设计,分析了控制系统中各组件的结构关系,重点设计了系统总线、中央处理器;软件部分分为打开文件、更新文件、监测运行、数据连接四步;为了检测系统的可行性,与传统并行处理控制系统进行实验对比,结果显示,设计的并行处理控制系统能够充足的利用系统CPU,精确地处理存储器中易失性用户大数据;该系统具有超强的工作能力,值得推广使用。

关键词:存储器;易失性用户;大数据;并行处理;并行处理控制系统

Control System for Parallel Processing of Volatile User Data in Memory

Liang Xiao[1,2]
(1. Qinghai Police Vocational College, Xining 810000, China; 2. Hohai University, Changzhou 213000, China)

Abstarct: Traditional parallel processing control systems have low CPU utilization when dealing with large data of volatile users in memory, resulting in poor processing and control precision. In order to solve this problem, a new large data parallel processing control system is designed. The hardware and software of the system are designed separately. The structure relationship of each component in the control system is analyzed, and the system bus and the central processing unit are designed emphatically. The software part is divided intofour steps: opening files, updating files, monitoring operation and data join. In order to test the feasibility of the system, the experimental results show that the designed parallel processing control system can make full use of the system CPU and accurately process the large data of volatile users in the memory. The system has super working ability and is worthy of popularization and application.

Keywords: memory; volatile user; big data; parallel processing; parallel processing control system

0 引言

随着科学技术的发展,互联网技术得到了迅速发展,“大数据”这个词汇逐渐流行起来,大数据技术逐渐渗透到经济、社会、交通等各个方面,为人们的生产生活掀起了一场数据革命,大数据技术已经成为评价一个国家综合国力的重要组成部分[1]。随着各行业领域科学技术的不断提升,互联网上升和累计的用户大数据规模日益庞大,网络数据爆炸式的增长,大数据时代已经来临[2]。然而现有一般技术难以对大数据进行处理,海量非结构化数据在存储、传输、分析及快速处理都存在一些问题。

并行处理是一种计算机处理的关键技术,是计算机在进行大量数据的实时采集、控制和处理时,为缩短延迟时间,提高计算机速度,实现高速处理的一种手段,即利用时间上的重叠或资源上的重叠来提高并行度[3]。计算机对数据的处理是以二进制代码形式进行的,这就使得人们不能直观的感知数据的传输与存储,使数据具有很强的隐蔽性。计算机数据若未被故意篡改,很少受影响而发生变化,数据具有很强的客观性[4]。但是计算机数据是动态存在的,很容易被外界窃取、修改、销毁,可能会随着时间的推移而改变,导致数据具有很强的脆弱易失性[5]。

本文在存储器中易失性用户大数据并行处理控制系统的硬件和软件框架进行设计,根据所设计框架阐述了易失性用户大数据并行处理控制系统的工作流程,通过实验验证了存储器中易失性用户大数据并行处理控制系统的可行性。

1 易失性用户大数据并行处理控制系统硬件设计

储存器中数据一般有易失性数据、网络数据、硬盘数据3种,如同用户登入、网络连接、运行进程等具有强时态性的数据信息,就属于易失性数据的范畴[6]。易失性数据具有隐蔽性、客观性、脆弱易失性的特点。易失性用户数据包含用户所用计算机基本配置信息、当前系统的时间记录、当前系统运行进程列表、登录用户近期活动记录、启动文件和剪贴板中的数据[7]。

并行处理是以同时采用多个处理单元处理输入信息的方式来达到缩短执行任务所花费时间的目的,是一种提高处理速度的最有效的技术,具有很大的提升空间。加速比与处理单元个数和任务并行度关系密切,任务并行度会制约并行处理系统的性能,在任务并行度为确定值时,增加处理单元个数而获得的加速比会达到一个极限值。

传统上的并行处理控制系统是通过单机操作系统设定

收稿日期:2018-09-10; 修回日期:2018-09-25。

作者简介:梁 霄(1984-),男,山西人,讲师,主要从事计算机应用方向的研究。

分时模式从而实现多任务管理的。并行处理控制系统一旦进入运行模式，任意节点都能够对共享数据进行更新。系统会将原有数据局部或全部作废，同时，数据更新采用消息点播的方式通知其它参与该操作的节点[8]。在并行处理控制系统的整个操作过程中，数据更新工作要确保在系统更新过程中不会出现资源活锁。并行处理控制技术包括多处理器紧耦合技术、共享存贮器技术、直接存贮器技术、高速缓冲存贮器技术、流水线处理器等[9]。

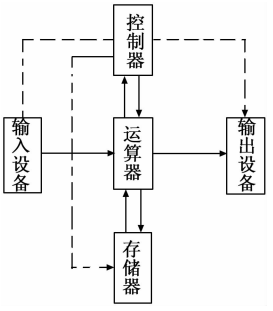

图 1 易失性用户大数据并行处理控制系统结构关系

如图 1 所示，易失性用户大数据并行处理控制系统结构的中心是运算器，输入设备和输出设备要经过运算器与存储器进行数据传送，不同硬件设施之间是串行工作的。当系统中的指令和数据通过运算器进行运算后，存储器再按地址访问的顺序将数据依次进行线性编址，也就是每当执行完指令后，进行下一条指令的执行。虚拟存储技术能够大幅度提高存储空间容量。数据并行处理控制系统的使用能够节省大量解决问题的时间，存储器中的易失性用户大数据可以随时进行读写，当电源关闭时，易失性用户大数据不能被保留，若想被保留，则需要将它们写入长久存储器中。

存储器中易失性用户大数据并行处理控制系统硬件设计具有两大特性：1）计算机内部信息流动的方式是以指令作为驱动；2）计算机主要有数值计算和数据处理两个方面的应用。图 2 为存储器中易失性用户大数据并行处理控制系统硬件设计，它的硬件结构较为复杂，具体情况如下：

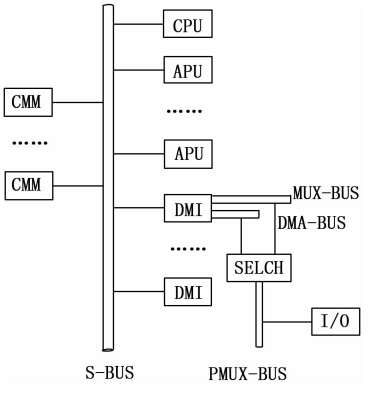

图 2 易失性用户大数据并行处理控制系统硬件设计

如图 2 所示，存储器中易失性用户大数据并行处理控制系统硬件设计中，S－BUS 为系统总线，带宽为 64 MB/s，是各个系统组件访问主存的高速数据通道，用于多个处理器之间的通信服务。S－BUS 为整个系统提供控制功能，并且把优先权分为支援高、循环、简单三级别，优先权越高越能被优先执行，这样就保证了越关键的信息越被首先处理，提高系统效率。并且当没有高优先权任务时，系统总线 S－BUS 允许其他任务占用，这样也就保证了所有任务共享 S－BUS，增加系统的吞吐量。

CPU 为中央处理器，APU 为辅助处理器，CMM 为组合存贮器，DMI 为直接存贮器，SELCH 为通道选择器。CPU、APU、CMM、DMI 都接在系统总线 S－BUS 上，其中，设定一个处理器为 CPU，其他处理器为 APU，最多设定 8 个 APU，不同处理器完成不同的功能。CPU 负责系统控制和执行计算工作，APU 专门负责执行计算工作。系统每个处理器都有权访问执行队伍，能执行自己任务的调度能力，CPU 负责维护执行队伍，能有效增加系统的处理效率。当某个 APU 出现故障时，其他 APU 会自动承担它的工作，是整个系统正常运行。当 CPU 出现故障时，系统会停机并发出报警信息，随即通过控制诊断系统进行诊断，有管理员决定重新设定一台 APU 临时执行 CPU 职能，维持系统正常运行。这样保证了系统的安全性和可靠性。

CMM 支持多道交叉存取，其内存为 16 MB，最高可以达到 1024 MB 系统内存，每个 CMM 都可进行线路的错误检测并进行纠正电路操作。当控制器进行诊断、记录和地址分配时，纠正电路会自动进行，以保证系统安全稳定运行。

DMI 直接连接在系统总线 S－BUS 上，以适配器的角色运行于输入设备、输出设备和 S－BUS 之间。DMI 为 S－BUS 进行识别和服务，对外部设备发出 DMA 请求，并收集和放置 I/O 命令。DMI 会生成 MUX－BUS 和 DMA－BUS 两条互相独立的子总线，其中，MUX－BUS 为多路转换总线，用于处理器控制下的编程传输；DMA－BUS 为直接存贮器存取总线，允许 I/O 控制器对存贮器进行存取。整个系统最多可支持 5 个 DMI，使多个数据模块并行传递。

SELCH 负责 MUX－BUS 和 DMA－BUS 两条子总线间的选择转换。SELCH 具有闲置和活动两种状态。当 SELCH 处于活动状态时，I/O 控制器与内存直接通过 DMA－BUS 交换数据，最高传输率为 10 MB/s。当 SELCH 处于闲置状态时，MUX－BUS 进行数据传送。PMUX－BUS 为私有多路传输总路，是 SELCH 生成的最下面一级的总线。各类 I/O 控制器都连接在 PMUX－BUS 上。各个子总线和系统总线以及各个设备在存储器内形成直接通道，使数据交换可以独立于处理器进行并行操作的。

2 易失性用户大数据并行处理控制系统软件设计

易失性数据指的是当计算机系统断电后不再存在的数据，通常保存在内存或硬盘临时文件中。收集易失性数据需要坚持以下几个原则：首先，要保证数据在采集过程中的原始性；其次，要保证数据在分析和传递过程中的完整

性；再次，数据从最初获取到被提取，要保证数据的连续性。易失性用户大数据并行处理控制系统最大的优势在于性能，系统具有高度可扩展性、高度容错性、高性能、较低分析延迟、易用且开放接口的特征。系统横向大规模可扩展，能满足数据量爆炸式增长，当查询失败时，只需重做部分工作，能够更好地维系系统的运行，具有很强的适应能力。

在数据存储器的并行查询和处理中，数据分布策略是关键问题之一，直接影响系统运行效率。大数据进行存储与并行处理时，需要在不同节点上对同一组数据进行同时处理，形成多个作业，这就使得系统产生明显的热点区域，从而造成严重的资源损耗量过大的问题[10]。大数据并行处理控制系统中各个位置产生的概率均等，因此，大数据并行处理可以被均衡分布在每一个节点中。数据在系统中的处理不是以文件为单位，而是以道集为单位进行的，能够有效地提高易失性用户大数据的读写速度，存储器中易失性用户大数据并行处理控制系统软件设计如图 3 所示。

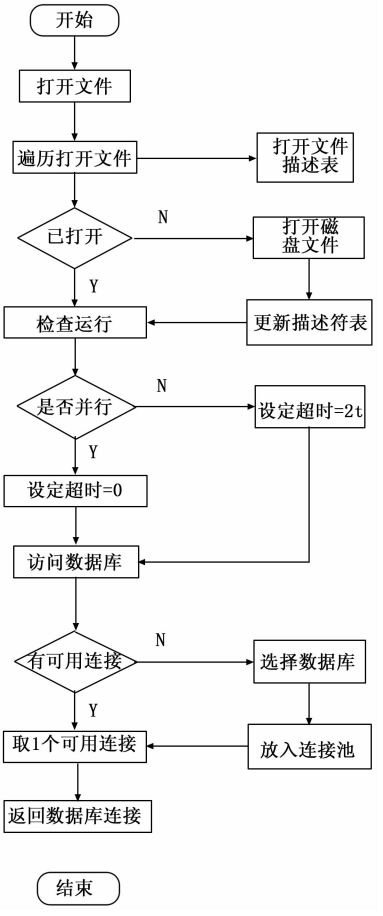

图 3　易失性用户大数据并行处理控制系统软件设计

文件描述表对一个进程中所有打开的文件列表进行实时记录，还可以对文件进行超时状态查询。频繁对文件进行操作会增加存储器负载，使系统程序运行效率降低，因此，维护文件描述符表运行尤为重要。如图 3 所示，系统开始后，首先遍历打开文件，查阅正要打开的文件是否已经被打开。若文件没有打开，则需要调用系统，打开磁盘文件，然后更新描述符表。若文件已经打开，则进行超时时间设置。

超时时间设置可以避免并行处理控制进程之间产生竞争条件。系统检查运行是否并行处理，若未进行并行处理，则进行超时设定，设定超时时间为串行时间 t 的 2 倍，若系统并行处理，则设定超时时间为 0，结束超时时间设置。运用超时机制的设定，有效地阻止多个进程同时占用要打开的同一文件，造成存储器长时间阻塞现象的发生，能够有效维护系统其他进程的正常运行。在系统运行时，并行状态的竞争几率出现较低，因此，并行设定优于串行设定。

大数据时代的数据量远超过单机数据量容量，数据分析是大数据处理的核心，大数据价值的完整体现需要多种技术的协同。为了便于对存储器中易失性用户大数据进行管理，需要在文件系统之上建立数据库。在程序运行时，传统操作中连接数据库过程需要花费很大一部分时间，这很容易造成服务器的过载。数据库连接池是指将可用数据库连接存储于一个缓冲池中，数据库连接池的使用，可以有效提高多用户访问数据库的效率。完成超时设定后，将进行数据库连接池进行数据处理。当用户程序需要连接数据库前，首先要查看数据库连接池中是否有可用的连接，如果有，则直接取一个可用连接，如果没有，系统将连接数据库，返回数据库连接对象，放入连接池。用户在选择数据库时，首先验证客户端数据库连接池，这样程序可以节省网络资源和运行时间，最大程度上提高网络利用效率。

3　实验研究

为了检测本文在存储器中的易失性用户大数据并行处理控制系统的实际工作效果，与传统的大数据并行处理控制系统进行对比，设定实验参数，将传统的大数据并行处理控制系统和本文在存储器中的易失性用户大数据并行处理控制系统进行比较，分别记录系统的运行状态和系统精密度测试，根据两个实验结果分析两种大数据并行处理控制系统的工作效果。

3.1　实验参数设置

在实验中，电源类型为高效节能电源；显示器分辨率像素数为 1600 * 1280，平均扫描速率为 64 KHz；系统 CPU 主频为 2.7 GHz，CPU 硬盘容量为 1T，转速为 7200 rpm；存储器内存容量为 16 GB；系统工作电压为 220 V，工作电流为 50 A，工作功率为 200 W；设备接口类型为 I/O 型，工作环境温度为－20～70 ℃。

3.2　实验结果与分析

得到的实验结果如下所示。

(1) 系统的运行状态结果。

观察图 4，将在存储器中的传统的大数据并行处理控制系统和本文的易失性用户大数据并行处理控制系统的运行状态，从系统的运行速度、CPU 利用率、内存使用率、网络流量和 I/O 读写率几个方面进行比较。传统的大数据并

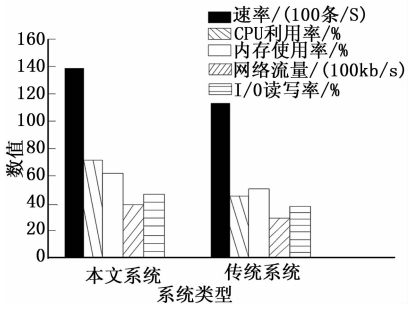

图 4 系统运行状态结果

行处理控制系统运行速度为 11 300 条/秒，本文中的易失性用户大数据并行处理控制系统运行速度为 14 000 条/秒。传统的大数据并行处理控制系统的 CPU 利用率为 45%，本文中的易失性用户大数据并行处理控制系统的 CPU 利用率为 71%。传统的大数据并行处理控制系统的内存使用率为 51%，本文中的易失性用户大数据并行处理控制系统的内存使用率为 62%。传统的大数据并行处理控制系统的网络流量为 3000 KB/s，本文中的易失性用户大数据并行处理控制系统的网络流量为 4000 KB/s。传统的大数据并行处理控制系统的 I/O 读写率为 38%，本文中的易失性用户大数据并行处理控制系统的 I/O 读写率为 47%。本文存储器中的易失性用户大数据并行处理控制系统的运行速度、CPU 利用率、内存使用率、网络流量和 I/O 读写率都高于传统的大数据并行处理控制系统，也就是说，本文存储器中的易失性用户大数据并行处理控制系统性能优于传统的大数据并行处理控制系统。

(2) 系统精密度测试结果。

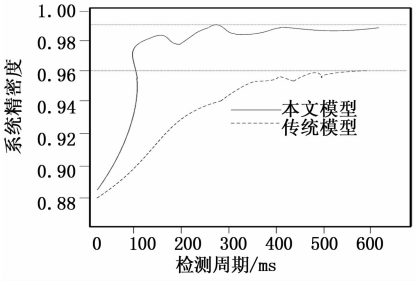

图 5 系统精密度测试结果

由图 5 可知，随着易失性用户大数据并行处理控制系统数据并行处理检测周期的增加，系统检测精密度会有所增加。传统的大数据并行处理控制系统大约经过 400 ms，系统检测精密度达到一个比较稳定的值，值约为 0.96。本文中的易失性用户大数据并行处理控制系统大约要经过 250 ms，系统检测精密度就能达到一个较为平稳的值，值约为 0.99。与传统的大数据并行处理控制系统相比，本文中建立的易失性用户大数据并行处理控制系统，大大缩短了系统检测所需的时间，提高了系统的工作效率，与此同时，系统数据处理的精密度也得到了一定的提高。

传统的大数据并行处理控制系统和本文在存储器中的易失性用户大数据并行处理控制系统在一定程度上都能够缩短延迟时间，提高计算机数据处理的速度。但是本文中的在存储器的易失性用户大数据并行处理控制系统性能明显优于传统的大数据并行处理控制系统。并且，该易失性用户大数据并行处理控制系统数据处理精密度更高，工作效率更高。

综上所述，本文在存储器中的易失性用户大数据并行处理控制系统性能好、工作效率高、结果稳定性强、精密度高，能够有效降低人工劳动强度，具有很好的发展潜力。

4 结束语

随着物联网、云计算、社交网络等新兴服务的发展，人类社会数据的规模和种类正高速增长，数据开始从作为简单的处理对象转变为一种基础资源。大数据的规模效应给数据管理、分析、存储带来巨大的挑战。大数据是一个较为抽象的概念，表示数据规模的庞大。大数据具有规模性、高速性、价值性和多样性的特点。大数据的数据来源广泛，但最基础的处理流程一致，将大数据进行并行处理，采用批量处理的方式，最大限度的利用系统存储器的内存，使大数据处理的工作效率得到有效提高。

本文研究的在存储器中的易失性用户大数据并行处理控制系统虽然具备一系列优点，缺少一定的实际操作基础，在未来的使用中可能存在一些潜在问题，并且系统对异构硬件的支持有限，会影响其扩展性，系统日常运维成本较高，仍具有较大的发展空间，需要进一步研究和探讨。

参考文献：

[1] 王子硕，翟岩龙，陶文俊，等. 高通量仿真中数据存储与访问技术研究 [J]. 系统仿真学报，2017，29 (9)：2016－2024.

[2] 邢德奇，康 乐. 大数据技术在北斗用户管理系统的现状分析 [J]. 电子技术与软件工程，2017，13 (14)：198－198.

[3] 朱永利，李 莉，宋亚奇，等. ODPS 平台下的电力设备监测大数据存储与并行处理方法 [J]. 电工技术学报，2017，32 (9)：199－210.

[4] 潘 巍，李战怀，杜洪涛，等. 新型非易失存储环境下事务型数据管理技术研究 [J]. 软件学报，2017，28 (1)：59－83.

[5] 张 倩. 论电力用户侧大数据分析与并行负荷预测 [J]. 纳税，2017，13 (29)：165－165.

[6] 王晨曦，吕 方，崔慧敏，等. 面向大数据处理的基于 Spark 的异质内存编程框架 [J]. 计算机研究与发展，2018，55 (2)：246－264.

[7] 秦 静，钱雪忠，王卫涛，等. 一种处理不平衡大数据的并行随机森林算法 [J]. 微电子学与计算机，2017，34 (4)：22－27.

[8] 赵 曦. 基于仿生算法的显著性特征数据挖掘方法 [J]. 西安工程大学学报，2017，31 (2)：244－250.

[9] 邢德奇，康 乐. 大数据技术在北斗用户管理系统的现状分析 [J]. 电子技术与软件工程，2017，66 (14)：198－198.

[10] 王小燕，张丽敏. 基于大数据的数据挖掘引擎研究 [J]. 电子设计工程，2017，25 (15)：31－34.