文章编号:1673-8411 (2017) 03-0114-03

地面气象观测数据入库多线程并行设计与实现

曾行吉, 李莹, 宋瑶
(广西气象信息中心, 广西 南宁 530022)

摘要:通过对地面自动气象站观测数据入库的工作流程分析,结果表明多线程处理性能比单线程高 3.85 倍,业务应用证明了该并行方案稳定可靠,为气象资料处理并行化规划和提高处理性能提供参考。

关键词:气象观测数据;入库;多线程;并行;设计
中图分类号:TP31 **文献标志码**:A

Parallel Design and Implementation of Meteorological Observation Data Storage Base on Multi Thread Processing

Zeng Xing-ji, Li Ying, Song Yao
(Guangxi Meteorological Information Centre, Nanning Guangxi 530022)

Abstract: Based on the work flow of ground observation data warehousing of automatic weather stations, the results show that performance of multithreaded processing is 3.85 times higher than the single thread and business applications show that the parallel scheme is stable and reliable. This can provide the reference for parallel programming and meteorological data processing.

Key words: meteorological observation data; storage; multithreading; parallel; design

各类气象探测仪器和各级气象业务工作人员按气象业务标准形成了大量的气象观测数据。气象观测数据具有数据格式多样、文件多、频次高、数据量大、时效要求严等特点。气象观测数据实时汇集到各级数据中心,进行数据打包与分发、格式检查、解码入库、数据分析、产品生成、文档制作等处理[1-3]。当前气象观测数据种类与数据量激增, 气象观测数据处理普遍要求实时完成处理。提高海量数据处理性能,使气象数据实时到达用户桌面,有利于提升气象服务响应速度。目前,多核处理器在计算机系统占主导地位,它在单个芯片内部集成了多个计算单元,同一时刻支持多条指令并发执行,而且采用超线程、缓存等技术加快指令执行速度。同时, 多核处理器通常采用共享式主存结构, 可以提供更快的同步操作与更高的核间通信带宽[4]。如果要充分发挥多核处理器的性能,必须采用多线程或多进程来执行,使得每个计算单元在同一时刻都有线程在执行。实践证明[5-10],利用多线程并行处理数据较串行处理可取得明显的性能提升。本文针对气象观测数据实时处理的高要求,以地面自动气象站观测数据为数据源,讨论气象地面气象观测数据解码入库并行性, 设计气象数据处理并行化方案并实现。该并行化办法可应用于其它气象资料多线程并行处理, 提高气象资料处理性能。

1 地面自动气象站观测数据

目前,广西全区地面自动气象观测站已有 3300 多个站,分为国家级站和区域站两类,其中区域站细分为 5 要素站、温雨 2 要素站、单雨量站等。另外还有船泊气象观测站、交通气象观测站等自动气象观测站。数据格式包括长 Z 格式地面自动气象观测报文、短 Z 格式地面自动气象观测报文、交通自动气

收稿日期:2016-10-14
基金项目:广西气象局 CIMISS 系统业务支撑能力建设创新团队
作者简介:曾行吉(1980-),男,工程师,硕士研究生,主要从事气象资料处理与质量控制工作。E-mail:269044875@qq.com

象观测站报文等数据格式,数据格式多样。不同站的观测要素不同,其中国家站观测要素最全,达 100 多个要素,数据项目多。每个站每 5 分钟上传一次报文到省数据中心，观测时间点后 2 分钟内省数据中心报文到报率超过 90%,数据洪峰突显。在多种因子影响下地面自动气象站观测数据会产生异常[11],部分异常会导致气象观测数据报文格式错误，尤其是在天气过程时,异常数据出现的概率会升高。

2　地面自动气象站观测数据入库处理流程

地面自动气象站观测数据入库处理实质是数据形式变换,由文本数据转变为数据库记录,其解码入库的流程为:FTP 远程获取报文→从本地缓存目录读取报文内容→解码→连接数据库→数据入库→记录日志→删除或备份缓存文件。同时,数据流程中可能出现的文件异常、格式异常、入库后报文删除等情况[12-13],如图 1。

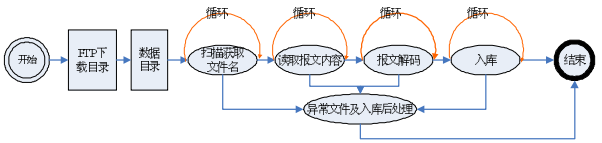

图 1　地面自动气象站数据解码入库流程图

3　地面自动气象站数据解码入库并行化方案

并行规划的关键是识别并发后合理规划并行,要尽可能避免资源竞用。从地面自动气象站数据解码入库工作流程分析,FTP 下载、文件扫描、文件读取、报文解码、数据入库、异常处理、入库后处理的环节都可以并行。但由于自动气象站数据解码入库工作流程与数据流具有明显的工作步骤：获取报文→读取报文内容→解码→连接数据库→数据入库→记录日志，将各工作步骤处理用工作流的方式组织可有效减少并发控制,有利有提高并行度,提升性能。工作流按文件划分处理单元，分打包文件、单站报文、状态报文、辐射报文、日数据报文、交通自动站报文。每类文件处理过程设计为一个工作流。报文分别进入相对应类别的工作流中进行处理，走过一个工作流程后，也就完成了该文件中的自动气象站资料解码入库工作。FTP 数据下载到本地目录工作独立性强，必然与数据解码入库并行。目录扫描是为了提取目录中报文文件,作为解码入库的数据源,考虑到目录扫描的频率不能过高,以 10 秒时间间隔扫描一次为宜,不适合与报文读取、报文解码、数据入库等数据处理环节一起并行，因此目录扫描也独立为一并行过程。

4　气象观测资料并行处理实现

由于资料入库时效要求高,1 分种内所有数据入库完毕;支持大数据量处理,不能产生数据堆积,在文件数可达数万到数十万个时，软件要求保持稳定运行。因此必须引入并行处理机制。

4.1　并行编程模型

不同的并行计算机体系结构往往对应不同的并行编程模型。一般来说并行编程模型存在两种划分:按照交互方式划分和按照并行方式划分。按照交互方式划分可分为基于共享内存的编程模型和基于消息传递的编程模型，基于共享内存的编程模型有 Windows 线程库、pThread、OpenMP、CUDA 和 Linda 等;基于消息传递的编程模型有 MPI、PVM 和 Pregels 等。按照并行方式划分可分为基于任务并行的编程模型和基于数据并行的编程模型，基于任务并行的编程模型有 Cilk、TBB 和 X10 等，基于数据并行的编程模型有 MapReduce、Dryad 和 Piecol 等。实际工作中,常根据硬件宿主特性的选择混合编程模型,常用的混合并行编程模型有:MPI+0penMP,CUDA+OpenMP,CUDA+MPI 等[14-16]。

4.2　气象观测资料并行实现技术

考虑到省级气象资料处理业务软件多是单机运行,运行平台是多核 Windows 服务器加,由于多核平台属于共享存储体系结构，所以多核编程本质上属于共享内存的编程模型,典型的有 Windows 线程库、pThread、OpenMP、CUDA 和 Linda 等[8-10],为充分利用 Windows 操作系统的优势,选用 Windows 线程库支持的多线程技术进行并行开发。多线程实现并行具有如下优点:(1)线程的创建和上下文切换的开销小;(2)线程间通信的方式多且简单高效;(3)多线程有庞大的基础库作为支持;(4)多线程的程序比多进程的程序更容易理解和修改。

根据地面自动气象站数据解码入库并行化方案,用线程封装工作流,线程每执行一趟即走完一个工作流,完成一批数据 FTP 下载或一个文件解码入库工作。为减少线程管理工作量，将线程分为 FTP 线程、目录扫描线程、人机交互线程和报文入库处理

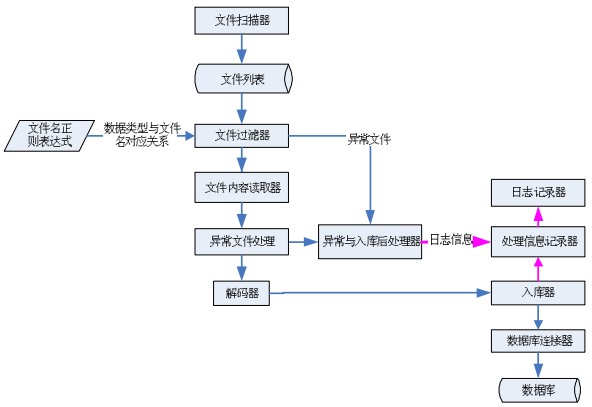

图 2 地面自动气象站数据解码入库软件设计框架图

线程 4 类。在软件运行初始化时,激活 1 条 FTP 线程,1 条目录扫描线程,8 条报文入库处理线程,人机交互线程利用 WinForm 异步机制实现。

自动气象站资料入库业务流程就是接收数据后解码入库过程,按工作流形式规划数据流动环节,清晰地划分各组件。如图 2。

地面自动气象站数据解码入库软件主要包括目录文件扫描、文件过滤、报文内容读取、报文解码(支持的报文包括:打包文件、观测数据报文、日数据报文、状态数据报文、辐射数据报文、交通站观测数据报文)、数据入库、数据库连接与操作、异常处理、入库后处理、日志登记、FTP 数据下载等模块。

5 应用情况

在双核 2.8G CPU、4G 内存、7K 转硬盘、Windows7SP1 环境下,用 10000 个自动气象站单站报文文件分别测试单线程和多线程处理速度:多线程用时 95s,单线程用时 366s,多线程是单线程处理性能的 3.85 倍。

地面自动气象站数据解码入库软件全区业务使用,极大地提高了区气象信息中心自动气象站资料处理业务能力,在自动气象站数据到达省中心后,1 分钟内完成解码入库目标,保证数据实时到达用户桌面,很好地满足省级气象预报、气象服务的时效要求;同时解决了广西全区 14 个自动站分中心站资料服务时效严重滞后问题,取得了很好的业务效益。

6 结论

地面自动气象站数据的数据量大、格式多样、观测要素多,偶然出现报文异常、入库实时要求高,非常有必要并行处理,文中以工作流思想设计了地面自动气象站数据解码入库并行方案,并用 Windows 线程库的多线程技术实现,测试表明性能较单线程提升了 3.85 倍,业务应用证明了该并行方案稳定可靠,为数据并行规则提供参考。总之,采用并化处理技术可有效提高数据处理性能,是海量气象资料加工处理有效的手段之一。

参考文献:

[1] 詹利群,黄祎萱,陈德诚.自动气象站中心站资料传输流程优化实践[J].气象研究与应用,2013,34(3):68-71.
[2] 王丽玫,任晓炜,李涛.广西气象信息网络传输业务实时监控系统的设计和实现[J].气象研究与应用,2011,32(S2):273-274.
[3] 魏莉.报文传输业务中的常见问题及解决方法[J].气象研究与应用,2007,28(S3):111-112.
[4] 高蕾,赖明澈,龚正虎.面向多核处理器的多实例并行 BGP 协议模型设计与实现[J].计算机工程与科学,2011,33(7):12-17.
[5] 杨尚琴,许自龙,洪承煜.基于多线程的地震相干体属性提取算法[J].计算机系统应用,2012,21(11):72-75.
[6] 吴石磊,安虹,李小强,等.组网雷达估测降水系统并行化方案的设计与实现[J].计算机科学,2012,32(3):271-275.
[7] 刘青昆,滕人达,刘凤,等.多核并行技术在分子动力学模拟中的应用[J].计算机工程与设计,2011,32(10):3395-3398.
[8] 王文鼐,陈邦文,韩鹏,等.采用多线程并行调度的网络仿真加速[J].南京邮电大学学报,2015,35(1):33-37.
[9] 何宗斌,张宫,樊鹤,等.一种基于并行计算技术提高测井数据处理速度的方法[J].石油天然气学报,2012,34(7):76-79.
[10] 梁启君,梁军,黄骞.雷达回波外推算法的并行化及实现[J].地理空间信息,2013,11(6):47-50.
[11] 徐明芳.CAWS600 型自动气象站定时数据异常的故障处理[J].气象研究与应用,2007,28(S1):103-105.
[12] 唐兵兵,杨帆,廖伟平,等.广西气象报文编发常见错误及处理方法[J].气象研究与应用,2010,31(3):104-107.
[13] 许嘉玲,王超球,赵秀英.自动气象站数据异常的原因分析[J].气象研究与应用,2007,28(S2):190.
[14] 林英,张雁.基于多核架构的软件开发方法研究[J].现代计算机,2013,(1):17-21.
[15] 蔡佳佳,李名世,郑锋.多核微机基于 OpenMP 的并行计算[J].计算机技术与发展,2017,17(10):87-91.
[16] 于方.多核多线程并行编程模型研究及应用[J].阴山学刊,2012,26(2):30-33.