

文章编号:1673-8411 (2018) 02-0094-04

# 基于多线程流水线的光盘自动刻录技术研究

张金标, 张恩红

(广东省气象探测数据中心, 广州 510641)

**摘要:**海量气象数据的长期归档保存是气象局的重要业务之一,传统的数据归档方式已经不能满足需要。基于多线程流水线的气象数据光盘自动归档方法可以解决海量数据归档的问题。利用该方法针对这一问题提出很好的解决方案,并完成了系统的开发、获得较好应用效果。系统通过基于线程流水线技术的多任务并行处理机制,大大缩短了海量数据的分类、压缩、光盘数据生成与刻录的整体运行时间。该系统每天能处理 981.8GB,刻录数据 577.3GB(以 50G 的蓝光光盘为例)。

**关键词:**海量数据;多线程;流水线;并行处理;自动归档

中图分类号:P49

文献标识码:A

## Research of CD automatic recording technology based on multithreading pipeline

Zhang Jinbiao, Zhang Enhong

(Guangdong Meteorological Survey Data Center, Guangzhou Guangdong 510641)

**Abstract:** Long term archival storage of mass meteorological data is one of the important businesses of the meteorological bureau. The traditional way of data archiving has been unable to meet the needs. The meteorological data compact disc auto archiving system gives out a perfect solution for this problem, which has completed system development and gained good application results. This system integrates data processing, disc burning, cover printing, and disc management, which solved the long term storage, auto archiving, and management problems of massive meteorological data. The whole processing time of massive data classifying, compressing, disc data organizing and burning is drastically reduced by using the multitask parallel mechanism based on multithreading pipeline technology. The system can handle 981.8GB data and record 577.3GB data per day (take the 50G blue light disc as an example).

**Keywords:** massive data; multi-thread; pipeline; parallel processing; auto archiving

## 0 引言

气象数据是气象业务运行的根本,它为天气预报、气候分析、灾害防范提供完备的资料,在气象现代化发展过程中起着十分重要的作用。数据中心负责全省气象观测数据的采集、分发、共享;负责本地区的气象产品的制作、分发、共享;负责国家局下发的数据的数据的分发和共享。数据中心每天收集本地观测数据、本地业务单位输出数据及国家局下发的卫星、

数值预报、全国观测数据超过 500G。要将这些数据刻录光盘并进行长期归档保存,是非常繁重一项任务。由于在刻录之前需要对这些数据进行分类、处理,在归档之后要求能够快速检索到所需的信息,如果仅通过人工来处理,将耗费大量的人力物力。目前,对单一数据的处理和归档的研究较多<sup>[1-3]</sup>,而对海量电子数据的归档处理研究主要是利用高效的存储设备<sup>[4-7]</sup>或数据库<sup>[8,9]</sup>方面的技术来提高存储效率和存储容量,并不特别注重软件的数据处理能力,难

收稿日期:2017-12-05

作者简介:张金标(1968-),男,高级工程师,理学学士,主要从事气象信息化和应用气象工作。

以适应本项目复杂的数据处理方式和较高的实时数据吞吐要求。因此,研制一套完善的、稳定的、高效的气象数据资料自动归档系统是非常有必要的,也是迫在眉睫的。

气象数据归档具有数据量大、小文件多、数据处理复杂、实时吞吐量大等特点,系统需要进行数据复制、分类、加密、压缩、建立索引、数据库存储、生成刻录数据、光盘刻录等诸多耗时操作,单任务单进程的处理系统将无法满足刻录系统的要求。线程是 CPU 可调度的、比进程更小的程序片段,线程的创建和线程间的调度的开销远远小于进程的开销,所以,将每个刻录任务作为一个线程,将多个线程并行处理,可充分利用当前计算机系统的多核 CPU 优势,提高运行效率。但是,多线程存在资源竞争问题<sup>[12,15]</sup>,例如,当多个线程并行执行大量文件复制时,由于存在对硬盘 I/O 接口的竞争,读写效率受到很大影响。因此,如果让各个线程并行操作不同的运算部件,资源

竞争将大大减少,有利于提高系统数据吞吐量。因此,我们采用线程流水线技术<sup>[10-15]</sup>,将刻录任务合理的分为多个步骤,使每个步骤尽量操作不同硬件资源,然后,每个步骤作为一个线程,同一任务的多个线程顺序执行,后一个任务的某个步骤必须等待前一个任务的相同步骤执行完成后才能启动。这样既可以提高系统的处理能力,也可以最大限度地解决 CPU、各运算部件、硬件设备 I/O 接口的资源竞争问题,全面提升系统性能。

## 1 方案设计

### 1.1 架构设计及组成

该方案采用 C/S 架构,由自动刻录系统、光盘信息管理系统、容灾系统、远程刻录监控系统四个部分组成,如图 1 所示。本文主要讨论自动刻录系统的功能实现,实线框内的部分,这也是该系统的核心部分。

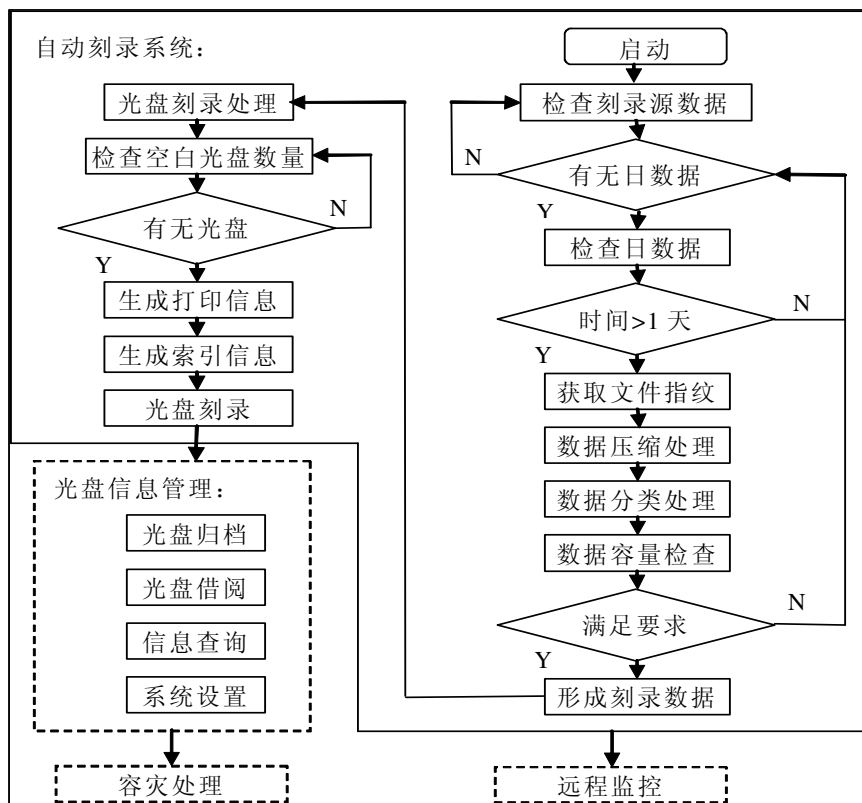


图 1 系统组成及流程图

### 1.2 光盘自动刻录系统的功能设计

该系统(图 2)是在整个大系统中的一部分,属于刻录管理下面的自动刻录设置模块下,通过这些配置,自动刻录系统自动完成整个刻录系统的数据

处理、光盘刻录、盘面打印等功能。本文讨论的是该系统的核心内容,主要分为以下六个模块:数据处理,包括文件目录分析、生成 MD5 数字指纹、数据库存储;数据压缩,包括文件分类设置、压缩属性设置、

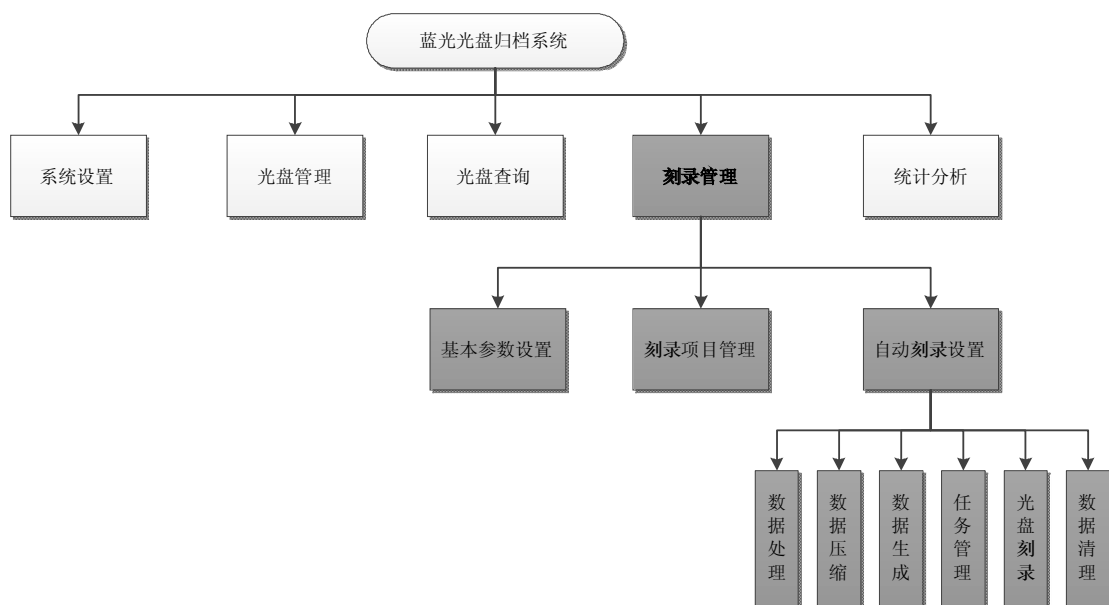


图2 光盘自动刻录系统

目录压缩;数据生成,包括光盘刻录信息(光盘说明及文件信息索引)、光盘刻录数据(分类数据文件、压缩文件、二维码图像文件、数据索引文件、工具文件等);刻录任务管理,包括生成刻录任务、刻录任务检查、提交刻录任务、双机刻录处理;光盘刻录,包括刻录状态检查、刻录错误处理、刻录完成处理、刻录校验、盘面打印;数据清理,光盘刻录完成信息、待清理数据信息、清理数据。

在系统实现时,把该六个模块独立封装,系统统一调度。由于这六个模块所实现的功能所消耗的资源也不同,因此具备了多线程流水作业提高效率的可行性。数据处理主要是从远程目录,抓取数据到本地,同时分析所抓取的文件,并生成相应的 MD5 码,这个环节主要是消耗 IO 和 CPU 资源;数据压缩和数据生成两个模块主要是消耗内存资源;刻录任务管理和光盘刻录主要是要把数据推送到光盘刻录机,对光盘刻录系统的监控,消耗的主要是 IO;数据清理负责清理磁盘中已经刻录的旧数据,主要是消耗 CPU。如果这些过程在一个进程线性完成的话,大概需要 7、8 个小时,每天只能完成三张光盘的刻录,不能完成业务的要求。为了提高效率,把这个过程分成 6 个部分,由 6 个不同的线程来完成,再加上不同作业可以流水线似的提交,整个系统的效率就大大提高了,每天可以刻录近十张光盘,基本可以达到业务的需求。

## 2 系统实现

### 2.1 数据库设计

气象观测的数据文件数量极为庞大,每年提交刻录的各种文件至少上亿个,这些文件的信息都要保存到数据库中以备以后查询,对于如此庞大的数据,显然用传统的数据处理方式无法完成。

该系统通过以下方式解决数据处理查询问题,首先基于分层结构的分块查找,即按照项目、光盘集、年度日期、目录、文件等逐层访问的方式建立数据块,这样可以极大地缩小数据查询的范围,同时大大减少数据查询的总量;其次,建立聚集索引,提高数据查询速度,同时在数据建立的过程中,数据的插入按聚集索引进行顺序插入,减少由于中间插入造成数据库处理性能的降低;最后,建立数据汇总索引,大量的操作通过汇总数据进行处理,减少了对最终文件的访问频率,极大提高了数据执行的效率。

### 2.2 自动刻录

自动刻录系统自动完成整个刻录系统的数据获取、数据处理、光盘刻录、盘面打印等功能。该系统采用 Windows Service 后台服务程序,保证 24 小时不间断运行和断电重启后在非登录状态下自动启动。

气象数据具有数据量大、小文件数多,系统需要进行数据处理、文件压缩、生成刻录数据、光盘刻录等诸多耗时操作,因此,我们将任务的每个步骤作为一个线程,同一任务的多个线程顺序执行,后一个任

务的某个步骤必须等待前一个任务的相同步骤执行完成后才能启动,形成多线程流水线执行结构,以最大限度地利用 CPU、各种运算部件、硬件设备和 I/O 接口,解决资源竞争问题,全面提升系统性能。

### 2.3 多线程流水线技术

流水线处理模式就是根据数据流向,将一个业务模块产生的数据元组通过数据管道直接传递给使用它的下一个业务处理模块。流水线上的业务处理模块处于运作状态,直到流入流水线的的数据被处理完毕。将流水线上的每一个业务模块对应一个单独的线程,就可以得到系统运行的多线程流水线模式<sup>[11]</sup>。

流水线结构的主要特点是系统不会由于线程数的增加而使处理能力下降,而是保持在一个恒定的水平上<sup>[10]</sup>。这一点类似于 QUEUE 方式的事务处理模式,当处理需求增加时,暂时不能处理的请求被放置在队列中,而不会像分时系统中那样直接在系统中分配处理资源,而对系统的性能造成影响。采用多线程流水线技术,可以合理的利用系统资源进行并行处理,充分的协调了系统中高速设备与低速设备之间的冲突,大大的提高了系统的吞吐量。

### 2.4 基于多线程流水线的并行自动刻录模式

本系统中的刻录任务包括数据处理(包括文件目录分析、MD5 数字指纹生成、数据库写入等)、数据压缩、生成光盘刻录数据(文件按光盘定义分类复制存放)、刻录任务管理、光盘刻录(刻录机读取指定目录数据到刻录机内存并执行刻录)、数据清理等六个步骤。这六个步骤是顺序执行的,执行时使用不同的 I/O 部件或对硬件的操作不同,所以可以采用六级流水结构并行处理各个刻录步骤,提高执行效率。

将每一个任务表示为  $T_1, T_2, T_3, \dots, T_i, \dots, T_n$ , 第  $i$  个任务  $T_i$  的六个步骤  $T_{i1}, T_{i2}, T_{i3}, T_{i4}, T_{i5}, T_{i6}$  分别作为一个线程顺次执行。当第  $i$  个任务的第  $j$  个线程  $T_{ij}$  进行数据处理时,采用线程同步技术以保证与前一个任务  $T_{i-1}$  的同级线程  $T_{i-1,j}$  不同时执行,以避免线程间的资源竞争。这样,最多可以同时并行处理 6 个刻录任务,充分利用了多处理器 CPU 的处理优势,减少了资源竞争,提高了数据处理的速度。

采用六级多线程流水线模式后,7 个刻录任务所用时间相当于 2 个刻录任务单独执行所用时间,执行速度提高了 3.5 倍。容易推算,当任务更多时,六级多线程流水线模式的执行加速比<sup>[8,9]</sup>最大将接近于 6。

## 3 系统功能测试

### 3.1 系统配置

处理器采用 Intel Xeon E5 2609 V3 6 核 CPU, 双 CPU 配置,共有 12 个处理器,内存 32GB,硬盘 3 \* 2T,做了 Raid 5 磁盘阵列,4T 硬盘空间可用。

### 3.2 并行处理测试

单线程测试:只有几个处理器能够得到较好的利用,CPU 的利用率在 10%~25%之间。

多任务并行处理测试:12 个处理器都得到较好的利用,CPU 的利用率在 90%以上。

### 3.3 压力测试

测试内容:对 2017 年 05 月 07 日至 2017 年 07 月 03 日间的数据进行统计,检测结果是否满足大于 500G/天的处理能力。刻录任务管理是对刻录系统和刻录任务的错误检查,占用时间忽略不计,不需要进行性能测试。

数据来源:2017 年 05 月 07 日至 2017 年 07 月 03 日间的气象系统实时观测数据。

测试结论:通过对各个执行环节处理能力的压力测试,系统满足数据处理能力不小于 500GB/天的要求。

## 4 结束语

本系统采用基于线程流水线技术并行处理海量气象数据,对它进行分类、压缩、光盘数据生成与刻录等耗时操作,可以充分利用 CPU、各种运算部件和硬件设备,解决资源竞争问题,同时,通过数据库的分层结构设计提高数据库处理能力,通过蓝光技术提高大数据光盘刻录速度,全面提升了系统运行效率。在此基础上,设计并开发了集数据处理、光盘刻录、盘面打印、光盘管理、远程监控和容灾恢复于一体的海量气象数据归档管理系统,系统各项功能和指标均达到业务要求,自动化程度高,运行稳定可靠。

### 参考文献:

- [1] 荣裕良,张霞,马忠芬,等.松江智慧气象为农服务系统开发研究[J].气象研究与应用,2017,38(1):102-106.
- [2] 陈少斌,苏彦,蒙绍臻,等.基于 Google Maps API 技术的自动气象站数据查询系统[J].气象研究与应用,2014,35(2):93-95.
- [3] 周展程,孙志强.基于 GIS 的自动气象站数据业务平台的

(下转第 100 页)



汽较大的云中进入干燥层后,温度传感器上的水或冰蒸发或结冰,与大气进行热交换,测量结果也会产生误差,雨水直接打在湿度片上也会影响水汽测量的准确度。留存记录的方法跟气球下沉记录一样,通过打印或是屏幕截图进行存档。

### 4.3 平滑曲线

在进行数据处理之前,先平滑曲线,在放大 10 倍的状态下及时删除明显的飞点。

### 4.4 终止层和终止时间确定

在确定了终止层和时间后,删掉球炸后的记录,保存数据,再关闭软件。

## 5 结束语

高空气象探测工作会遇到各种不同的情况,我们要在业务工作中不断总结经验,加强学习,对发生的事故原因和处理方法及时总结,掌握各种情况的处理方法和技巧,才能保证观测工作的顺利进行。以上是我们在工作中遇到的情况及处理方法和技巧,

供同行们参考。

### 参考文献:

- [1] 中国气象局主编.常规高空气象观测业务规范[M].北京:气象出版社,2010:27-28.
- [2] 张春阳,彭力.1981-2010 年洪家气象站风向风速变化特征分析[J].气象研究与应用,2017,38(2):72-76.
- [3] 杨玉静,杨志华,农国傲,等.桂平市 20 年大风天气气候统计分析[J].气象研究与应用,2017,38(1):47-49.
- [4] 何如,谭敏玲,罗红磊,等.南宁市横县地区风能资源评估[J].气象研究与应用,2015,36(3):59-63.
- [5] 杨凤群,林苗青,翁永安.南澳岛风能资源的评估[J].气象研究与应用,2011,32(3):58-61.
- [6] 孙小龙,苏志,何如等.富川地区风能资源分析[J].气象研究与应用,2012,33(3):25-28.
- [7] 崔丽曼.河南省雷暴大风气候特征及近年变化趋势分析[J].气象研究与应用,2016,37(3):33-37.
- [8] 李刚,钱保贵,刘清芳等.滇东南高空风季节变化特征分析[J].气象研究与应用,2012,33(2):68-70.
- .....
- (上接第 97 页)
- 实现[J].气象研究与应用,2016,37(1):91-93.
- [4] 严文瑞,曹强,姚杰,等.一种面向大容量光盘库的新型文件系统[J].计算机研究与发展,2015,52:1-8.
- [5] 缪嘉嘉,付印金,余沛毅,等.低能耗磁光混合归档系统的设计与实现[J].计算机技术与发展,2017,27(8):52-56.
- [6] 赵伟东.电子档案蓝光存储应用探究[J].档案学研究,2015,3:88-95.
- [7] 杨鹏.基于蓝光存储技术的气象电子档案存储管理[J].信息与电脑,2015,(16):84-86.
- [8] 王瑞通,李炜春.大数据基础存储系统技术研究[J].计算机技术与发展,2017,27(08):66-72.
- [9] 宋丽娜.海量存储系统中数据分级存储的关键技术研究[D].长沙:国防科学技术大学硕士论文,2011.
- [10] 曾行吉,李莹,宋瑶.地面气象观测数据入库多线程并行设计与实现[J].气象研究与应用,2017,38(3):114-116.
- [11] 黎颖智,史彩霞,刘世学.关于异步多线程快速提取 CIMISS 数据入库方法的研究与应用[J].气象研究与应用,2018,39(1):118-120.
- [12] 胡晓峰,张亮红,刘文怡,等.VC 多线程流水线在数据存储系统中的设计与实现[J].电子器件,2016,39(4):964-967.
- [13] 季圆圆,张涛,王海鹏.基于线程流水线的多核线程调度策略[J].计算机工程,2013,39(2):279-287.
- [14] 王晶,樊晓桢,张盛兵,等.多核多线程结构线程调度策略研究[J].计算机科学,2007,34(9):256-258.
- [15] 曹斌,张霞,刘积仁.多线程 Web 服务器设计中的流水线技术[J].东北大学学报(自然科学版),1999,20(3),229-231.
- [16] 刘云生,陈世俊.基于多线程的流水线查询处理方式[J].计算机应用,2004,24(6):54-56.