文章编号:1671-4598(2019)12-0047-05 DOI:10.16526/j.cnki.11—4762/tp.2019.12.011 中图分类号:TP399 文献标识码:A

基于分布式的玻璃缺陷检测技术研究及性能优化

孟 陆,金 永
(中北大学 信息与通信工程学院,太原 030051)

摘要:玻璃检测速度的的提高会在短时间会产生大量图像数据,传统分布式框架 MapReduce 处理速度和及时性无法满足玻璃缺陷检测的要求;课题将 MapReduce 分布式框架运用到海量图像处理,设计阈值分割算法完成对玻璃缺陷图像的处理;通过添加数据划分模块使计算与存储本地化,加快数据处理的及时性;实验结果表明改进的 MapReduce 计算框架处理速度平均提高 14.1%,能够对运行速度为 600 m/h 的玻璃带进行在线检测,并检测出玻璃带上缺陷的个数、位置和缺陷的类型。

关键词:缺陷检测;分布式系统;图像分割;数据本地化

Research and Performance Optimization Based on Distributed Glass Defect Detection Technology

Meng Lu, Jin Yong
(School of Information and Communication Engineering, North China University, Taiyuan 030051, China)

Abstract: With the increase of glass detection speed, some defects of MapReduce distributed computing framework are exposed, and the processing speed and timeliness cannot meet the requirements of industrial glass defect detection technology. Based on the MapReduce parallel computing framework, the paper designs a threshold segmentation method to complete the segmentation of glass defect images. By adding a streaming data processing module and a data partitioning module, the computing and storage are localized, and the timeliness of data processing is accelerated. The experiment results show that the improved MapReduce computing framework has an average processing speed increase of 14.1%. It can detect the glass ribbon running at 600 m/h and detect the number, position and type of defects on the glass ribbon.

Keywords: defect detection; distributed systems; image segmentation; data localization

0 引言

随着数字图像检测技术在生产领域的广泛应用,很多应用需要及时地分析当前的生产状况,传统的数字图像检测系统难以满足工业生产的需求。以玻璃生产加工业为例,在原料加工、制备、熔化、澄清和冷却等各种生产环节中,由于工艺制度的破坏或操作过程的差错,从退火窑出来的玻璃带往往带有不同类型和大小的缺陷。缺陷检测系统需要及时的控制横切机将包含缺陷过多、不符合国家标准规定的部分切除,从而达到提高整条玻璃生产线玻璃质量的效果。由于玻璃带传输速度的加快,缺陷检测系统短时间会采集大量高分辨率缺陷图像数据,要实现生产线上玻璃带缺陷的及时检测,需要采用与生产速度匹配的,及时不间断的在线检测系统[1]。

MapReduce 应用于大规模计算机集群处理海量数据的并行计算中,是一种基于键/值对的数据处理模型[2]。该模型将复杂的分布式计算过程分为 Map 与 Reduce 两个阶段。总任务提前被分割成若干个小任务,每个划分的小任务由一个 Map 任务来计算,Map 任务计算完成之后将中间结果传递给 Reduce 任务,进行全局的结果汇总并计算出最终的结果。

MapReduce 的出现在一定程度上缓解了大数据处理的难题。MapReduce 由于最初只定义了基于文本的数据类型,在默认设置中无法支持图像数据类型的处理。要实现对大量图像文件的分布式并行化处理,需要实现图像数据到 MapReduce 分布式计算框架所默认的数据类型转换。目前许多学者针对基于 MapReduce 实现图像并行化处理做了深入研究。文献[3]利用 MapReduce 所提供的读写接口来设计所需要的图像数据类型,由此来实现将图像序列化为 MapReduce 可进行处理的数据类型,达到图像并行化处理的效果。但这种方法存在的缺陷是当面临大量小文件存储的问题,会导致 Map 任务数量过多,造成系统处理效率低的问题。文献[4]提出了一种新型的图像数据并行处理模型。利用 MapReduce 模型适合处理大规模文本数据的特性,选择将存储了图像路径的文本文件代替图像数据进行输入,从而避免了设计图像数据类型的麻烦,但同样会面临大量小文件存储导致的存储效率低下的问题。

收稿日期:2019-05-28; **修回日期**:2019-06-17。

作者简介:孟 陆(1992-),男,硕士,主要从事光学信号处理方向的研究。

通讯作者:金 永(1977-),男,教授,主要从事图像处理、无损检测及机器视觉方向的研究。

其次，随着数据处理需求的提高，也暴露出 MapReduce 这种分布式计算框架一些缺陷。MapReduce 有其性能瓶颈：在 Map 和 Reduce 之间，隐形设有中间处理部分，比如为了让不同结果分发到对应的处理节点上，需要把所有结果汇总到每个节点上再进行排序，每个节点截取对应区间内的数据[5]。该过程是 MapReduce 之所以能够正确运算的关键所在，但是其影响了系统处理的速度。

为解决此问题，文献[6]提出本地增强负载均衡算法将 MapReduce 的流程扩展到，减少与 shuffle 重叠的计算并充分利用 CPU 和 I/O 资源，但开发难度较大且不易于扩展。文献[7]中提出了一种本地感知的 reduce 任务调度策略，考虑分区的位置和大小，采用默认的 hash partitioner 使 reduce 任务尽量本地化，以减少 shuffle 数据量，提高 MapReduce 的性能，但对大规模数据集性能提升不大。文献[8]提出基于数据本地化和负载均衡的任务分配策略，既减少了 Shuffle 阶段数据的传输量，同时也避免出现任务分配不均衡的情况。文献[9]中针对 shuffle 处理策略的不足，采取管道策略，将 map 生成的数据通过管道直接传输到 Reduce，降低了 I/O 代价，提高了效率，但缺少 shuffle 会导致计算准确率严重下降，不适用于大规模图像处理。

基于上述背景，本课题以 MapReduce 并行计算框架为研究基础，针对大量小文件的存储问题进行了存储结构的改进，通过添加流式数据处理模块和数据划分模块，使得计算与存储本地化，加快数据处理的实时性。并以在线所采集的大量玻璃图像为测试对象，通过改进 MapReduce 计算框架实现对各种玻璃缺陷及时准确的检测。

1 系统检测方案

对玻璃带进行缺陷的在线检测，文献[10]提出一种基于数字光栅投影的浮法玻璃缺陷检测方案，此方法利用位于玻璃带上方的高速 CCD 相机采集玻璃带表面呈现的图像，再由检测系统的对采集到的缺陷图像进行缺陷检测。由于单位时间内线阵 CCD 相机获得数据量十分巨大，目前无法采用单一的硬件和软件系统实现，由此本文设计了基于 Hadoop 集群的多路并行处理的玻璃带缺陷检测方案，如图 1 所示。

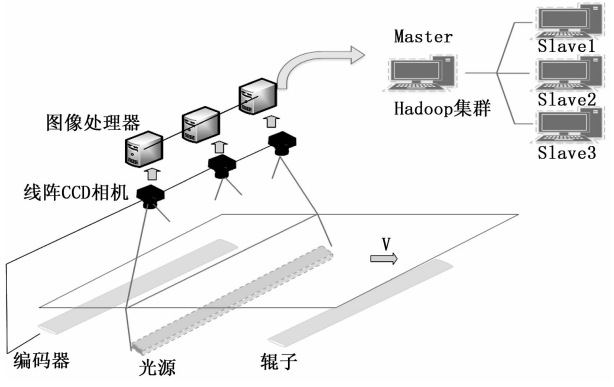

图 1 玻璃缺陷在线检测总体方案

该系统由光源、编码器、高速线阵 CCD 相机、图像处理器和 Hadoop 集群处理器组成。具体检测方法是由多路独立的图像采集单元和由 Hadoop 集群构成的图像处理单元组成，采用正透视的背光照射方式的检测光路[11]，通过多路高速线阵 CCD 相机并行采集玻璃表面的光强信号，整个 CCD 线阵都与 Hadoop 集群相连接。高速数据采集卡通过双 DMA 模式连续采集光强度信号，将其转换为灰度图像数据并传输到上位机。任务由客户端提交给资源管理器，CCD 摄像机收集的数据传输到 Hadoop 集群的 Master。Master 将每路 CCD 相机数据分配给对应的 Slave。然后，运行分配任务的 Slave 完成分割算法。最后，处理结果返回给 Master。一旦缺陷过多，由系统控制的横切机将切除具有缺陷的部分。

2 缺陷图像的分布式处理

2.1 图像存储结构的设计

在缺陷图像处理之前，首先要解决图像数据的存储问题。本文默认的数据块大小为 128M，玻璃缺陷图像数据大小远小于默认的数据块大小，直接存储会在 NameNode 中存储大量文件名称信息，从而严重降低集群的运算性能。同时访问大量小图像回频繁调用 I/O 接口，也会增加文件寻址时间。因此本文针对图像分片利用 HIPI 接口的 HipiImageBundle 类，将本地图片通过文件遍历方法上传，图像分片组成一个包含数据和索引的 hib 文件存储结构，生成 glass. hib 和 glass. hib. dat。hib 文件存储位移及索引信息，hib. dat 文件存储图像数据。通过对 HIPI 接口的调用，减少了大量图像分片对 Hadoop 性能的影响。

其次，由于 MapReduce 运算模型的性质，在图像处理过程中会将图像数据随机分块。如果不对图像数据进行预处理，而是由 Map 任务直接分块处理，在最后的聚合阶段无法将分块后的图像数据进行准确的聚合，从而无法还原到处理之前完整的原始图像。因此，在 Map 任务进行分块前，本文对玻璃缺陷图像进行预处理，提前将图像处理前的分块顺序和图像分块相对于原始图像的位置坐标存储在数据块的元信息中。在 Map 任务完成之后，可以将分块后的图像数据与提前存储的元信息进行比对，依照原信息中存储的数据块坐标位置即可快速完成分块图像的准确聚合。本文所采取的图像存储结构如图 2 所示。

2.2 MapReduce 功能的设计

MapReduce 缺陷检测实现过程可分为如下 3 个阶段：首先，图像序列中每个图像被分割成多个小的图像分片，并将图像分片分配到 Hadoop 的数据节点上。接着，数据节点上的每个 Map 过程独自完成图像分片的缺陷分割任务。最后，在 Reduce 过程中将检测后的图像分片聚合，获得最终检测结果。MapReduce 的工作流程如图 3 所示。

在 Map 阶段，通过 ImageInputFormat 接口读入图像分片，ImageRecordReader 函数负责输入以及对记录进行读取操作，得到分割记录和产生输入分片[12]。MapReduce 程序将输入的＜key，value＞对输送给 map 完成程序执行，采

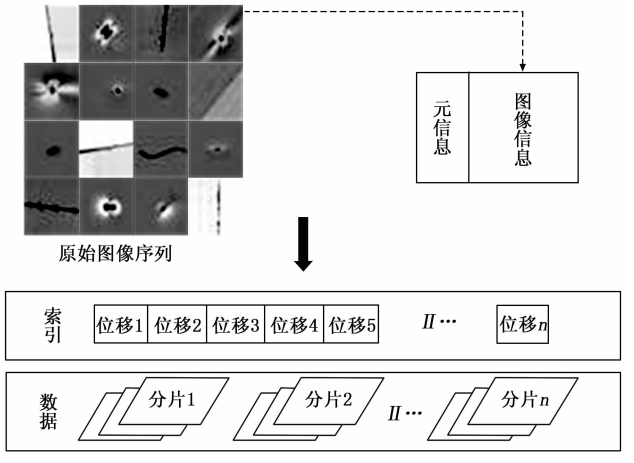

图 2　图像存储结构的设计

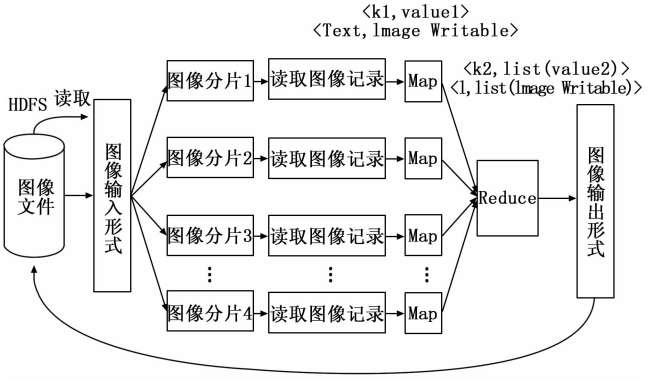

图 3　MapReduce 图像处理流程

用阈值分割的方法完成每块图像的缺陷检测。map 任务结束后，将检测结果以＜Text，image＞的键值对形式输出，结果发送到 reduce 任务，图像分片元信息中像素的索引号和坐标号保存在 key 中，value 则保存该图像分片的缺陷检测结果。

在 Reduce 阶段，针对 key 中保存的元数据信息和 value 中的检测结果，按照元数据中存储的索引号和像素坐标等信息，将检测结果归并，并将检测结果保存到 HDFS 中的不同文件夹中。在对图片调用执行完毕，再启动 reduce 程序把执行处理后图像进行合并操作。

3　缺陷检测算法优化

Map 阶段数据节点产生的中间数据需要经过网络传输到 Reduce 阶段的计算节点作为其输入数据，这个中间阶段称为 Shuffle[13]。Shuffle 阶段的数据传输和 Reduce 阶段数据存储非常耗时，所以如何减少 Map 阶段不必要的网络带宽占用，成为提升 MapReduce 作业执行效率的关键。而 Map 阶段性能与数据本地化相关，所以提升数据本地化可以有效提升 MapReduce 作业的执行效率。

Hadoop[14]数据本地化是指数据集无冗余地划分至每个节点，通过划分数据集来并行化数据的处理。如图 4 所示，在任务调度过程中，如果不考虑数据本地化，就会使得本可以直接从本地读取输入数据的任务需要跨机架通过网络访问来远程读取数据，增加了任务的执行时间。在改进的 MapReduce 计算模型中，数据本地化使系统能够感知机架。通过数据定位，数据分配可以提高系统的并行度，从而提高系统的处理效率。在本文中，数据块分区被提前到 Map 阶段，当 map 完成后，所有数据都被发送到相应的 Reduce 节点，然后进入 Shuffle 阶段，这个过程在简化繁琐的中间排序过程的同时也能很好保证运算的准确性，提高了传统 MapReduce 框架的效率。

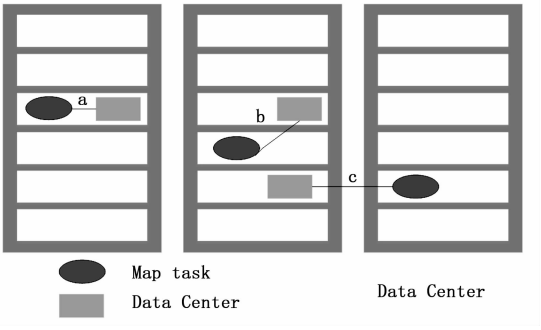

图 4　MapReduce 本地化机制

4　实验

本实验是在局域网内，实验搭建 Hadoop 集群由一个主节点和 4 个从节点组成。Hadoop 集群安装在虚拟机上，软件配置和硬件设置如表 1，表 2 所示。实验选取在线采集玻璃缺陷图像作为实验对象，缺陷图像的平均大小约 212KB。由于实验在分布式集群上进行，实验中数据块的副本数设置为 3，同时集群中默认的数据块大小为 128M，MapReduce 中 Reduce 的数目设置为 1，实验环境的软件和硬件配置如表 1 和表 2 所示。

表 1　实验环境软件配置

软件	版本
操作系统	CentOS 6.9
Hadoop	Hadoop 2.6.5
Java	JDK 1.8.0_191
Hipi	2.0
Ant	1.9.13
虚拟机	VMware Workstation 12 Pro

表 2　实验环境硬件配置

硬件	参数
内存	4 GB
硬盘空间	50 GB
网络	Network Address Translation(NAT)

在实验环境搭建完成之后，对在线所采集的玻璃缺陷图像进行基于 MapReduce 的分布式缺陷检测实验，结果如图 5 所示。图 5（a）是一幅含有疥瘤的玻璃缺陷图像，图 5

(b) 是在 MapReduce 上进行分块缺陷检测的中间结果，图 5 (c) 是将中间结果聚合后，得到的最终检测结果。

将原玻璃缺陷图像和检测后的结果进行对比，如图 6 所示，分别展现出了划痕、夹杂、污点和疥瘤的阈值分割结果。通过多次的测试可以看出，本文所采取的图像分割算法可以有效完成不同种类玻璃缺陷图像的分割。

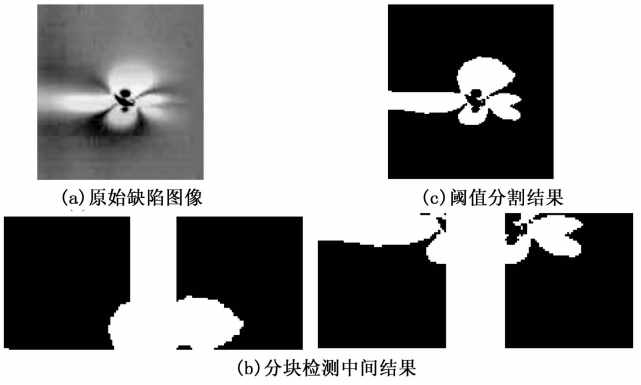

图 5 MapReduce 疥瘤缺陷检测结果

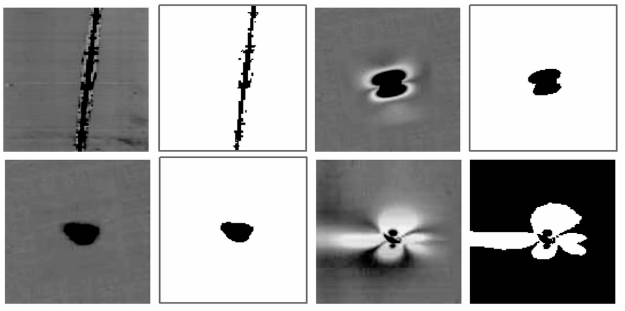

图 6 MapReduce 图像分割结果

为了验证本文改进的 MapReduce 框架处理效率和运算性能的提升，将改进的 MapReduce 框架与原来的运算框架分别运行在三节点和四节点的 Hadoop 集群，测试所采取的图像数据量分别在 200 M、400 M、600 M、800 M 和 1 000 M 的条件下检测两种框架运算的效率。通过图 7 和图 8 发现，不同节点 Hadoop 集群下，随着图像数据量的不断提高，处理时间显著加快，基本呈线性增长。对比三节点和四节点集群的处理时间，图 9 和图 10 分别展现出了不同节点下运算性能改进的加速比，从图 9 和图 10 的加速比变化可以看出，伴随着节点数的增多，MapReduce 的运行效率会有所提高，加速比大约从 1.1 提高到 1.24。集群节点数越多，加速比会适当提高，这体现出改进的算法在多节点集群上加速效果更明显，也可以说明本文改进的算法在数据调度和本地化数据方面改进显著。总之，改进的 MapReduce 架构改善了数据运算性能，因此本文所做的改进相对于传统的 MapReduc 框架在性能上有所提高，可以很好的应用于玻璃缺陷图像检测系统中。

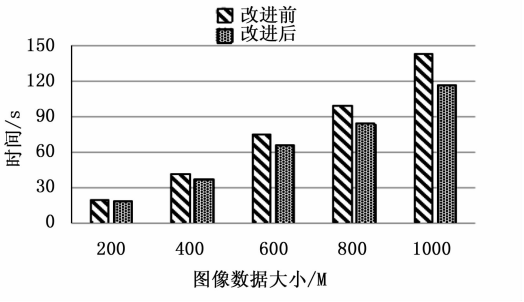

图 7 三节点 Hadoop 集群处理结果

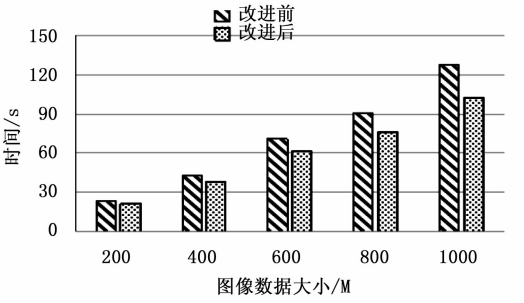

图 8 四节点 Hadoop 集群处理结果

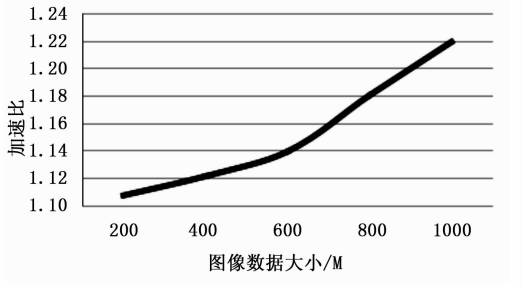

图 9 三节点 Hadoop 集群加速比

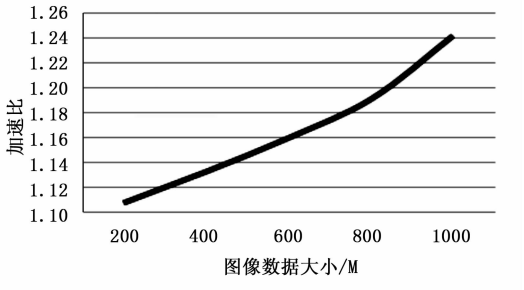

图 10 四节点 Hadoop 集群加速比

5 结论

本文针对目前玻璃检测系统无法满足及时性的问题，将 MapReduce 分布式计算框架应用于海量玻璃缺陷图像检测中。通过对存储结构的改进，解决大量小文件存储导致效率低下的问题，完成基于 MapReduce 的并行化玻璃缺陷图像阈值分割算法。此外，在原有的 MapReduce 计算框架基础上，对中间处理过程 Shuffle 做了进一步改进，通过数据本地化改善运算性能。实验表明，改进的 MapReduce 并行计算框架数据处理速度得到显著提高，保证了系统的准确性和及时性，为玻璃缺陷检测后续的打标和切割工序提供了有效信息。

参考文献：

[1] 余发山，田西方，韩超超，等. 玻璃生产缺陷在线检测技术研究 [J]. 河南理工大学学报（自然科学版），2013，32（4）：476－480.

[2] 张郝娟，吕晓琪，温秀梅，等. Hadoop 平台下基于内容的医学图像检索 [J]. 现代电子技术，2017，40（4）：115－119.

[3] 李　倩，施霞萍. 基于 Hadoop MapReduce 图像处理的数据类型设计 [J]. 软件导刊，2012，11（4）：182－183.

[4] 刘　军，李　威，吴梦婷，等. Hadoop 平台下新型图像并行处理模型设计 [J]. 计算机工程与应用，2019，55（6）：186－190.

[5] 熊　倩，张　奕，郭　明，等. MapReduceShuffle 性能改进 [J]. 计算机应用，2017，37（S1）：58－62，67.

[6] Li J J，Wang J，Lyu B，et al. An improved algorithm for optimizing MapReduce based on locality and overlapping [J]. Tsinghua Science and Technology，2018，23（6）：744－753.

[7] Hammoud，M，Sakr M F. Locality－aware reduce task scheduling for MapReduce [A]. 2011 IEEE Third International Conference on Cloud Computing Technology and Science（CloudCom）[C]. 2011.

[8] 王　浩. Hadoop 环境下基于数据本地化的 Reduce 任务调度策略 [J]. 计算机与现代化，2016（1）：114－120.

[9] Xie J，Tian Y，Yin S，et al. Adaptive Preshuffling in Hadoop clusters [J]. Procedia Computer Science，2013，18.

[10] 石兵华，金　永，王召巴，等. 基于数字光栅投影的浮法玻璃缺陷检测方法研究 [J]. 光电子·激光，2014，25（3）：521－525.

[11] 金　永，魏　博，王召巴，等. 基于双 CCFL 的玻璃缺陷检测技术研究 [J]. 中北大学学报（自然科学版），2013，34（1）：66－69.

[12] 翟锐涛. MapReduce 模型下的图像并行化处理研究 [D]. 西安：西安科技大学，2017.

[13] 田进华，张韧志. 基于 MapReduce 数字图像处理研究 [J]. 电子设计工程，2014，22（15）：93－95，100.

[14] 张　帅，贾如春. 基于 Hadoop 的大数据信息安全监控云平台设计与研究 [J]. 计算机测量与控制，2017，25（9）：72－74，78.

（上接第 41 页）

这就要求进行本测试的时间与位置尽可能的接近理想位置。

与其他测试方法相比，本测试所需要的设备更加简单，工程实施更加便捷，可行性较强，而且与标准相比达到的准确度也比较高。这证明我们的测试方法的确行之有效，且操作更为简便，对于 G/T 值测试的可实现性有了一定程度的提高。

地面站品质因数 G/T 值是衡量地面站接收性能的重要指标，是进行卫通系统链路；设计的重要依据，该测试方法可以解决工程上 G/T 值测试的需求，对于天线生产单位，系统设计单位的工程实施有着实际帮助。

本试验方法适用于 Ku 频段各型线极化卫星天线的测试，其他频段或极化方式的卫星天线的 G/T 值测试，使用本方法测试的效果，未经过作者验证。

5　结束语

本文阐述了载噪比比较法测量 G/T 值的原理方法和特点，利用标准喇叭、频谱仪、低噪声下变频器和功分器等几种简单器材，自行设计测试系统，来测试实际情况下的天线系统的增益 G 与接收系统噪声温度 T 比值 G/T 值，这个指标的准确测量可以衡量地面站灵敏度的质量。在操作流程简便，参试工具简单的条件下，本文设计的测试系统依旧测量出了与标准值相近的结果，相对于其他多种测量方式，本方法实行性更强，实用性更好。尤其在中国地处偏远、交通不便的条件下，测试仪器简单、实用性更强的系统更容易进行操作，从而节省成本，也能保证实际测试准确进行。

参考文献：

[1] 袁惠仁. 天线参数的射电天文测量 [M]. 北京：电子工业出版社，1986.

[2] 潘　捷. 卫星通信天线 G/T 值测量技术 [J]. 邮电设计技术，1993（3）：50－54.

[3] Qin S Y，Wang X Q. Accuracy considerations in the G/T value measurement of the earth station using carrier to noise ratio direct method [J]. Journal of China Institute of Communications，2000.

[4] 陈　辉，路志勇. 利用卫星源测量有源天线 G/T 值的简便方法 [A]. 全国微波毫米波会议 [C]. 2009.

[5] 李　文，马忠松. G/T 值测试三原则及 G/T 值测量方法 [J]. 国外电子测量技术，2011，30（6）：33－36.

[6] 毛乃宏，俱新德. 天线测量手册 [M]. 北京：国防工业出版社，1987.

[7] 夫顿纳基斯. 卫星通信手册 [M]. 成都电讯工程学院出版社，1987.

[8] 储钟圻. 数字卫星通信 [M]. 北京：机械工业出版社，2006.

[9] 林昌禄. 天线工程手册 [M]. 北京：电子工业出版社，2002.

[10] 殷　琪. 卫星通信系统测试 [M]. 北京：人民邮电出版社，1997.

[11] 秦顺友，许德森. 卫星通信地面站天线工程测量技术 [M]. 北京：人民邮电出版社，2006.

[12] 唐　波，梁兴东，李炎磊，等. 基于改进的扩展波数域算法的 SAR 实时成像方法 [J]. 国外电子测量技术，2010，29（7）：29－33.