

PDF 集群并行解析显示技术研究

罗明宇,付燕平,刘其军,归 强

(广东粤铁瀚阳科技有限公司,广东 广州 510630)

摘要:针对大型拼接显示系统对 PDF 文件高分辨显示的需求,文中研究了集群并行高分辨信息显示平台的显示技术,分析了 PDF 在大屏显示系统中的重要性以及 PDF 文件格式和层次关系,研究并探讨了 poppler 库和 mupdf 库的优缺点,以及对 PDF 文件格式的解析与显示技术,并基于 poppler 库和 mupdf 库分别实现了 PDF 文件的解析和集群并行显示。通过对比实验,采用 mupdf 库能够更清晰、高效地实现 PDF 集群并行显示,验证了文中提出的 PDF 集群并行解析显示技术可极大地提高大型拼接显示系统对 PDF 文件的高分辨显示处理能力。

关键词:PDF 解析;并行拼接显示;高分辨率显示

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2014)06-0243-04

doi:10.3969/j.issn.1673-629X.2014.06.061

Research on PDF Parsing and Parallel Display Technology on Cluster

LUO Ming-yu, FU Yan-ping, LIU Qi-jun, GUI Qiang

(Guangdong Railway & Sun Technology Co., Ltd, Guangzhou 510630, China)

Abstract: A PDF parsing and parallel display technology on cluster is presented to solve high resolution display problem. The importance of the PDF rendering in the tile high resolution system and the PDF format and hierarchical relationship are analyzed. The advantages and disadvantages of both poppler and mupdf library are studied to parse PDF information on cluster and parallel display on the tiled high resolution system. And use these two libraries to achieve the PDF file parsing and cluster parallel display on the tiled high resolution system. The comparison tests show that the display results are clearer and more efficient for PDF parsing with mupdf library in the cluster parallel tiled high resolution display system. Also, the display resolution of PDF information is greatly increased by the PDF parsing and parallel display technology.

Key words: PDF parsing; parallel tiled display; high resolution display

0 引言

随着计算机图形学的发展,信息量的不断增长,信息显示技术已经深入到了生产和生活的各个方面,例如指挥调度、工业设计、影视娱乐、科学研究、教育教学、安全管理、灾害管理、监控检测等。PDF 作为信息的载体,在网络信息记载、传递共享等方面起着重要的作用。

PDF 是由 Adobe 公司研发的电子文件格式^[1-2]。目前已经广泛应用于电子文档发行和数字化信息传播,PDF 正在成为数字化信息的工业标准。

研究表明,通过视觉人们可以在很短时间内对所见的信息进行快速有效地处理,因此将信息通过可视化方式快速显示在人们眼前显得尤为重要。尤其是对信息在大尺寸、超高分辨率、大信息量情况下的显示需

求越来越突出。提升信息显示系统的响应时间、处理速度和分辨率是信息显示技术一个重要的研究方向^[3-6]。在铁路指挥中心指挥调度系统中,采用并行集群技术构建的 SPIDer 超高分辨信息处理与显示平台集成了超高分辨率显示技术、显示拼接技术、多屏图像处理技术、信息显示同步技术、高速网络通信技术等,是指调度系统的信息处理、分析、管理和显示的综合应用平台。采用并行集群显示技术充分发挥了节点机的图形处理能力,通过在节点机上构建并行的信息处理显示功能,建立了高度协同的分布式集群处理平台,实现了信息的高分辨显示,系统具备了良好的灵活性、扩展性和高性价比等优势。目前,SPIDer 超高分辨信息处理与显示平台正广泛应用于铁路指挥控制系统的数据处理、路桥隧道健康监测、超高分辨率

收稿日期:2013-07-24

修回日期:2013-11-04

网络出版时间:2014-02-24

基金项目:广东省科技计划项目(2012A080102003);广东省省部产学研结合项目(2012B090500012)

作者简介:罗明宇(1971-),男,总工,博士,研究方向为通信技术、大数据处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0857.007.html>

GIS、卫星图片显示等工程和项目领域。

在指挥中心的显示系统中,通常借助计算机显卡 DVI 信号输出的方式,将 PDF 信息放大显示在指挥中心的超高分辨拼接显示屏幕上,这种显示方式的分辨率低,当显示窗口放大时,将不可避免地出现马赛克等情况,严重影响显示效果。为满足 PDF 在指挥中心的显示系统中超高分辨需求,文中研究了 PDF 集群并行解析显示技术,通过实验,该技术可极大地提高大型拼接显示系统对 PDF 信息的高分辨处理显示能力,为指挥决策提供清晰的信息显示。

1 SPIDer 超高分辨信息处理与显示

SPIDer 超高分辨信息处理与显示系统以 OpenGL 作为渲染引擎,利用 OSG 对场景信息进行管理,其系统结构包含一个主控制节点和若干个渲染子节点。SPIDer 系统由主节点来调度 SPIDer 各子节点机的渲染和控制,每台子节点机最多控制 4 块显示单元。主节点采用高性能的管理控制器,主要功能是负责对多显示区域进行坐标划分和时间调度管理。以数据流的方式控制子节点间的同步和子节点显示区域的划分,对子节点进行任务调度管理。SPIDer 的子节点机为显示单元建立了分布式的显示环境,并对场景进行绘制和裁剪,每台子节点机在主节点机的控制下进行并行数据处理,并对每块显示单元进行正确的渲染控制。子节点机和主节点机通过千兆以太网进行数据和控制流的交换。子节点机根据主节点机发出的指令对收到的信息进行分割,并对属于自己视景范围的信息进行渲染处理^[4]。每台子节点机只负责自己区域显示信息的解析和渲染。虽然每块显示单元的显示分辨率是有限的,但是多块显示单元拼接起来即可组成超高分辨

率的显示墙,SPIDer 为并行集群显示建立了超高分辨率的显示平台,提供了灵活、开放、可扩展的分布式高性能可视化显示系统,全面支持包括视频、文字、图像、地形、3D 等多种信息源的大型拼接屏幕高速、高分辨同步显示。

SPIDer 系统的信息显示是通过集群并行解析图形处理和并行绘制来实现的,与其他多屏显示系统^[5]采用集中处理、分发图元或者像素信息不同,SPIDer 是通过子节点机分布式独立处理解析数据、显示输出的,主节点机和子节点机间只有少量的消息同步、控制等通信,占用带宽小、响应速度快,因此,SPIDer 系统具有超强的并行处理能力和响应速度,支持信息的超高分辨显示。

2 PDF 文件格式

Adobe 公司提出 PDF 文件格式的目的是为了支持跨平台、多媒体集成信息的出版和发布,尤其是提供对网络信息发布的支持。为此,PDF 具有许多其他电子文档格式无法比拟的优点^[7-8],PDF 可以将文字、字型、格式、颜色及独立于设备和分辨率的图形图像等封装在单个文件中,该格式文件还可以包含超文本链接、声音和动态影像等电子信息,支持特长文件,信息集成度和安全可靠性非常高。

PDF 文件从根本上来讲是一个 8 字节的序列,PDF 文件格式和常用的 HTML、XML 等结构化文件格式基本上相同,都包含有关键字、分隔符、数据等。PDF 文件通常可分文件头(Catlog)、文件体、交叉引用表、文件尾(Trailer)等。PDF 文件的层次关系如图 1 所示。

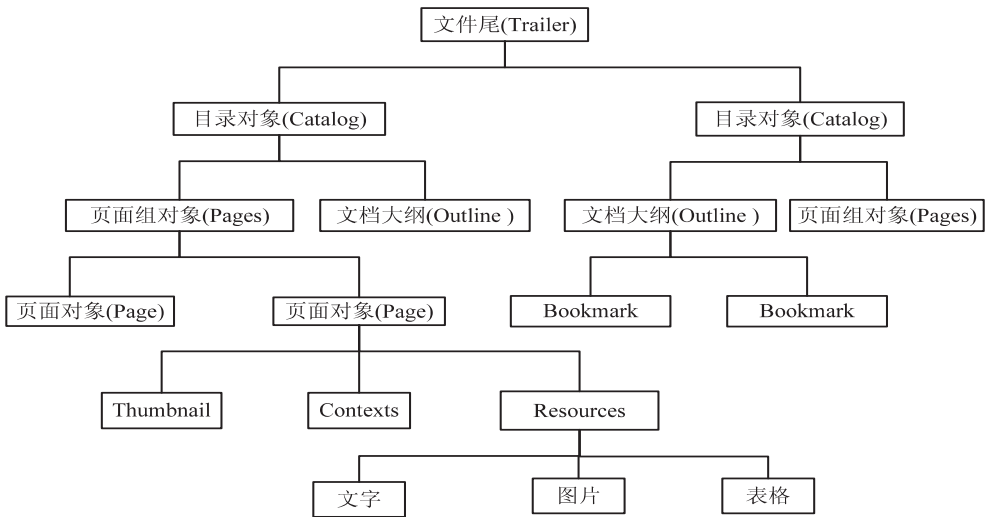


图 1 PDF 文件的层次关系

PDF 文件是由一些称为“对象”的模块组成的,并且每个对象都有数字标号,以便被其他对象所引

用^[9-10]。这些对象不需要按照顺序出现在 PDF 文件里面,出现的顺序是任意的,正是由于 PDF 文件页与

页之间的不相关性才保证了 PDF 文件页面可以随机访问。PDF 对象中的页面 (page) 是 PDF 文件中最重要的对象,包含如何显示该页面的信息,例如使用的字体、包含的内容(文字、图片等)、页面的大小等页面属性信息。这些信息是可以直接给出的,但是里面的子项更多的是对其他对象的引用,具体的信息存放在其他对象中。页面包含的信息是在一个称为流(stream)的对象里,这个流的长度(字节数)是直接给出或指向另外一个对象(包含一个整数值,表明这个流的长度)的。

3 PDF 集群并行解析与显示

目前,支持 PDF 解析读取的软件库比较多^[11-13],也有专门的阅读器支持 PDF 的解析读取操作,但是,在超高分辨并行集群显示环境下解析显示 PDF 信息,需要把 PDF 作为一个二维信号源来处理,进行独立的解析和渲染,可以采用 poppler 库或者 mupdf 库实现 PDF 信息的集群解析显示。

3.1 基于 poppler 库的 PDF 显示

poppler 是从 xpdf 继承而来的 C++类库,具有较全面的 PDF 文件交互操作功能。为了达到良好的输出效果,poppler 库需要与二维矢量图形绘制库 cairo 配合使用以实现 PDF 的解析,并且,poppler 利用 cairo 作为后端对 PDF 文件进行渲染,采用这两个库的渲染流程如图 2 所示。

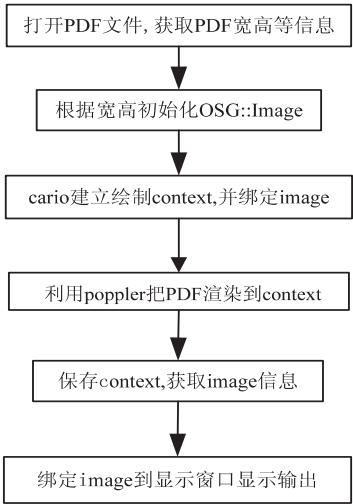


图2 poppler 和 cairo 渲染流程图

首先打开 PDF 文件获取 PDF 宽高等信息,然后初始化 OSQ::image,通过 cairo 建立 context 后,利用 poppler 渲染显示 PDF。

在测试实验中,poppler 和 cairo 对一些 PDF 格式渲染效果不理想,有时候 PDF 文件中有的 Object 不能很好地解析出来,读取的 PDF 渲染出来后会 出现白块像素丢失的现象,显示出来的 PDF 文件会出现白色矩

形,同时,在 SPIDer 超高分辨率显示系统上,PDF 显示会出现图像马赛克、边缘锯齿明显。

3.2 基于 mupdf 库的 PDF 显示

mupdf 是 Artifex Software Inc 研发的开源 PDF 读取库,通过简单配置,mupdf 就能够快速完成 PDF 的读取和解析任务。采用 mupdf 库实现 PDF 的集群并行解析,可集成利用 OpenGL 和 OSG 将其渲染输出显示,其渲染流程如图 3 所示。

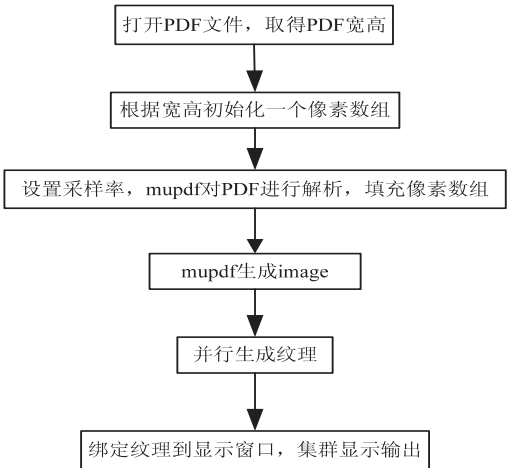


图3 mupdf 对 PDF 的渲染流程图

在获取了 PDF 宽高信息后,mupdf 库将对 PDF 进行解析,依据输出分辨率填充像素数组,然后生成 image 绑定纹理显示。值得一提的是,mupdf 支持对显示采样率的设置,通过调整采样率可实现对集群显示内容的放大,避免出现图像马赛克、边缘锯齿等情况。在实现的过程中利用多线程的方法同时对所有的 PDF 页进行解码并把解码出来的所有的 PDF 页分别生成的纹理对象存储到一个 TextureArray(纹理数组)中,每一页对应一个纹理单元。完成之后当进行纹理翻页时只需要从纹理数组中取出当前页的纹理进行渲染而不需要再次对当前页进行解码显示,这样就可以提高翻页和自动播放的效率。

在测试实验中,mupdf 库能够提供高质量的抗锯齿的图形解析渲染,支持多核、多线程的 PDF 文件解析,并以像素级别高精度、高品质地解析输出信息,支持以矢量显示以获得最高的显示保真度,在 SPIDer 超高分辨率显示系统上,高速实现 PDF 不失真的解析渲染显示。

虽然 poppler 的交互功能强于 mupdf,但考虑到 mupdf 库可以独立完成 PDF 解析渲染显示,无需 cairo 库配合工作,支持多线程处理,显示速度快,图像清晰,不出现像素丢失、锯齿等现象,符合指挥中心对显示系统的高分辨信息快速显示要求,因此,建议在集群并行显示系统中采用 mupdf 作为 PDF 文件的解析库。

4 对比实验

传统的拼接系统显示 PDF 一般采用投影拼接技术来实现,首先通过计算机解析 PDF 信息,然后从计算机显卡 VGA/DVI 输出口获取显示信号,交拼接显示系统投影输出。

由于计算机显卡的输出分辨率有限,当拼接显示系统的分辨率较高时,PDF 输出马赛克、锯齿等情况明显。其显示效果如图 4 所示。



图 4 PDF 显卡输出的显示效果

而基于 mupdf 的 PDF 集群并行解析显示的显示效果如图 5 所示。



图 5 PDF 集群并行显示的显示效果

通过以上显示效果对比可以看出,由于 PDF 显卡输出方式只以点阵形式输出,不支持矢量显示,所以图像放大后会出现马赛克、边缘锯齿等现象,影响显示效果。而 PDF 集群并行显示技术支持矢量显示,图像清晰,放大后不会出现马赛克,无边缘锯齿等现象。

相关对比实验的结果见表 1。

表 1 对比实验结果

对比项目	PDF 显卡输出显示	集群并行解析 PDF 显示
矢量信息显示	不支持	支持
细节显示	不支持	支持
显示分辨率	1 920×1 080 约 200 万像素	由显示单元数决定, 可达 5 亿像素以上

可见,PDF 显卡输出方式不支持矢量显示,显示分辨率低,难以满足高分辨率 PDF 显示的要求。而 PDF 集群并行解析支持矢量信息、细节信息等的显示,显示分辨率高。

另外,对于一些特殊的信息文件类型,例如 Word、PPT 等可以通过转化为 PDF 文件格式,实现超高分辨显示。

5 结束语

文中研究了 PDF 集群并行解析显示技术,通过分析 PDF 文件的文件头(Catlog)、文件体、交叉引用表、文件尾(Trailer)等文件格式和层次关系,研究并对比了采用 poppler 库和 mupdf 库的 PDF 集群解析显示技术。通过在集群并行高分辨信息显示平台的显示,与传统 PDF 拼接显示进行了实验对比,实验显示效果进一步证明 PDF 集群并行解析显示技术可支持矢量信息、细节信息等的显示,显示分辨率高,可为指挥中心等大型拼接显示系统的 PDF 集群并行显示提供有效的解决方案。

参考文献:

[1] Adobe Systems Incorporated. PDF reference; sixth edition [EB/OL]. [2010-10-23]. http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf.

[2] PDF specifications[S/OL]. 2011-08-08. <http://www.adobe.com/devnet/pdf/pdfreference.html>.

[3] 刘真,石教英,彭浩宇,等.基于 PC 集群并行图形绘制系统综述[J].系统仿真学报,2006,18(Sup):70-72.

[4] 孙益辉,陈福民.超大屏幕实时交互分布式渲染平台关键技术[J].计算机应用,2008,28(S1):230-231.

[5] DeFanti T, Leigh J, Renambot L, et al. The OptiPortal, a scalable visualization, storage, and computing interface device for the OptiPuter [J]. Future Generation Computer Systems, 2009,25(2):114-123.

[6] 石教英.分布式图形绘制技术及其应用[M].北京:科学出版社,2010:75-105.

[7] Bienz T, Cohn R, Meehan J R. Portable document format reference manual; version 1.2[S]. [s. l.]: Adobe Systems Incorporated, 1996.

[8] 李珍,田学东. PDF 文件信息的抽取与分析[J].计算机应用,2003,23(12):145-147.

[9] 陈俊林,张文德.基于 XSLT 的 PDF 论文元数据的优化抽取[J].现代图书情报技术,2007(2):18-23.

[10] William S L, David F B. Document analysis of PDF files: methods, results and implications[J]. Electronic Publishing-Origination Dissemination and Design, 1995,8(3):207-220.

[11] 李强,刘时进. PDF 阅读器的设计与实现[J].计算机工程与设计,2010,31(7):1635-1638.

[12] 杨道良.面向对象的中文 PDF 阅读器的设计与实现[J].计算机应用,1999,19(6):1-4.

[13] Cross J S, Munson M A. Deep PDF parsing to extract features for detecting embedded malware[R]. [s. l.]: Sandia National Laboratories, 2011.

PDF集群并行解析显示技术研究

作者：[罗明宇](#)，[付燕平](#)，[刘其军](#)，[归强](#)，[LUO Ming-yu](#)，[FU Yan-ping](#)，[LIU Qi-jun](#)，[GUI Qiang](#)
作者单位：[广东粤铁瀚阳科技有限公司, 广东 广州, 510630](#)
刊名：[计算机技术与发展](#) 
英文刊名：[Computer Technology and Development](#)
年，卷(期)：2014(6)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201406061.aspx