**数据压缩博士申请课题**

**The Design, Implementation, and Deployment of a System to Transparently**

**Compress Hundreds of Petabytes of Image Files For a File-Storage Service**

**一、主要工作**

我们报告了我们使用另一种技术的经验:基于统计模型的特定于格式的透明文件压缩，该模型经过调优，可以在大型语料库上表现良好。

We report on our experience with a different technique: format-specific transparent file compression, based on a statistical model tuned to perform well on a large corpus

它用一种定制的统计模型取代了最低层的基线JPEG图像——无损霍夫曼编码——我们对该模型进行了调整，使其在存储在Dropbox中的大量JPEG图像语库中表现良好。

called Lepton, that replaces the lowest layer of baseline JPEG images— the lossless Huffman coding—with a custom statistical model that we tuned to perform well across a broad corpus of JPEG images stored in Dropbox.

**二、实现思路**

以下是该部分的主要内容总结：

核心工具：Lepton作为一个独立的工具，可以对基线JPEG文件执行往返压缩和解压缩。它是开源软件，可以在多种操作系统上构建和运行。

特定设计约束：Lepton设计时考虑了分布式网络文件系统中实时压缩和解压缩的特定约束，包括跨独立块的分布、块内并行解码和流式处理。

JPEG压缩概述：JPEG图像文件由头部信息和图像数据（扫描）组成。Lepton使用现有的无损技术压缩头部，并使用算术编码压缩图像数据。

算术编码：Lepton用算术编码替换了JPEG文件中的Huffman编码，这是一种更高效的技术，可以更准确地预测系数值，从而实现更小的文件大小。

概率模型：Lepton使用一个复杂的自适应概率模型，该模型通过大量实际图像的测试开发而来。模型使用大量统计箱（bins），每个箱子跟踪特定上下文中“1”比特与“0”比特出现的概率。

上下文相关编码：Lepton的模型利用了图像中不同系数类型（如DC和AC系数）及其在块内的索引的上下文信息，以提高压缩效率。

并行解码和流式输出：Lepton能够将JPEG文件分割成独立的段，每个段由一个线程解码。这种并行化处理允许快速开始传输字节，即使整个块尚未完全解压缩。

Huffman交接词：为了支持多线程编码，Lepton在文件格式中包括了“Huffman交接词”，这允许解码器在文件中间或符号中间恢复状态。

内存分配：Lepton在处理前分配所有内存，以避免在处理输入数据时进行动态内存分配。

安全性：Lepton在读取输入数据前进入限制模式，只允许执行特定的系统调用，以增强安全性。

跨平台支持：Lepton支持Linux、MacOS、Windows、iOS、Android和Emscripten (JavaScript)平台。

部署策略：Lepton在Dropbox的生产环境中直接由后端文件服务器执行，或者在高负载时由专门的Lepton集群处理。

1. **实现难点**

**Round-trip transparency.**

Lepton需要确定地恢复原始文件的精确字节

Lepton needs to deterministically recover the exact bytes of the original file

**Distribution across independent chunks.**

Lepton必须能够解压缩JPEG文件的任何子字符串，而不需要访问其他子字符串。

Lepton must be able to decompress any substring of a JPEG file, without access to other substrings

**Low latency and streaming.**

为了实现这一点，Lepton格式包括“霍夫曼切换词”，使解码器成为多线程的，并在请求后很快开始传输字节。

To achieve this, the Lepton format includes “Huffman handover words” that enable the decoder to be multithreaded and to start transmitting bytes soon after a request.

**Security.**

**Memory.**

1. **图片压缩的常见方法**

通用熵压缩。

Generic entropy compression.

有损图像压缩。

Lossy image compression

格式感知像素精确重压缩。

Format-aware pixel-exact recompression

支持格式、保存文件的再压缩。

**Format-aware, file-preserving recompression**

最后一组工具可以通过往返恢复原始文件的精确字节来重新压缩JPEG文件。

A final set of tools can re-compress a JPEG file with roundtrip recovery of the exact bytes of the original file.