136

# 面向 IEEE COMTRADE 格式的海量录波数据 并行压缩 / 解压算法

桂勋

(电子科技大学 航空航天学院,四川 成都 610054)

摘要:提出了一种直接面向 IEEE COMTRADE 格式的海量故障录波数据并行压缩/解压算法。算法给出了 COMTRADE 数据文件中时间信息无损恢复公式,提出了针对状态量数据的优化 RLE 编码。对高频模拟量数 据采用提升格式小波变换,用硬阈值量化后采用基于 LZ77 和 Huffman 混合熵编码的 Deflate 算法进行压缩, 而对低频模拟量数据采用 RLE 编码。为有效分批压缩 COMTRADE 数据文件,提出了一种基于高频模拟量数 据最佳小波分解层数下的最少补零的数据划分算法。映射到不同 CPU 核心的压缩/解压线程通过竞争通道 序号获取计算数据。简述了压缩文件格式及其采用的变长存储和访问技术。试验证明此算法可获得大压缩 比,并且可随着 CPU 核心数的增加和 COMTRADE 文件的增加,获得线性加速比。

关键词:数据录波; IEEE COMTRADE;压缩;解压;数据处理;LZ77; Huffman; Deflate;小波变换 中图分类号:TM 73 文献标识码:A DOI: 10.3969/j.issn.1006-6047.2013.05.024

## 0 引言

电力系统电磁暂态过程是电力系统中短暂的、 但非常重要的物理过程。所有的故障都伴随着相应 的电磁暂态过程,现代电力系统故障诊断、广域保 护和稳定分析愈加依赖于电网动态监测系统采集的 含丰富故障信息的暂态过程信号。因此,对电磁暂 态的监测、记录和分析一直是电力系统中一个重要的 研究方向。近年来随着电子技术的快速发展.电力 系统内暂态录波逐步向高采样率、连续稳态记录和海 量存储方向发展印。为了提高故障录波数据的传输 效率和降低存储空间,众多文献提出了各种压缩算 法[2-15] 其中可分为直接离散小波变换压缩[2-7]、直接 熵编码[8-9]压缩以及这2种算法的综合[10-13]共3类。 其中在离散小波压缩研究领域,由于第1代离散小波 变换 Mallat 算法<sup>[14]</sup>相对于第 2 代基于提升格式的离 散小波变换算法[15]无法实现无损压缩,且计算量大、 需要额外内存[10-12,15],第2代基于提升格式的离散小 波变换和熵编码的综合压缩算法[10-12]成为当前此领 域内的研究热点。

然而到目前为止这些压缩算法在实际应用中却 存在以下问题。

a. 以往压缩算法是为了解决电话线传输速率过低的问题,但随着电力通信网络的改善,百兆、千兆以太网大规模安装,通信效率已经不再是难以克服的瓶颈,因此,目前众多厂家已经放弃了原有的复杂压

缩、解压方案,而直接在高速以太网上传输录波文件。 而另一方面,保信系统厂家不可能让负责上传录波文件的子站和主站软件系统为每个录波器厂家实现复 杂的压缩/解压通信协议。正因为这两方面的原因, 限制了原有压缩算法的实际应用范围。

b. 以往压缩算法是为了解决海量录波数据的存储问题。但在电力系统内,目前所有的暂态记录格式逐渐被统一到了 IEEE COMTRADE 1999<sup>[16]</sup>(下文简称 COMTRADE)标准上,而几乎所有的第三方暂态分析软件都是以 COMTRADE 标准为输入格式的。因此,即使在保信系统上采用了压缩/解压通信协议,在主站上还是要解压后以 COMTRADE 标准保存,否则第三方分析软件无法使用。因此压缩算法也没有解决实际中面临的海量 COMTRADE 数据存储问题。

c.随着多核和嵌入式多核的大规模普及,带来了 一场计算技术革命:传统的串行编程技术被完全颠覆, 逐渐被基于多线程模式的并行编程技术所取代。软 件要提高运行速度和效率已经不能依赖 CPU 主频, 而要使软件系统本身在设计上适应新的计算机体系 结构。而此领域内以往提出的算法都是串行算法, 已经无法适应计算机体系结构的发展。

笔者经过现场调研后认为,只有直接面向 COM-TRADE 标准文件的压缩算法才能彻底解决问题 a、 b,假如实现了通用的 COMTRADE 标准压缩/解压算 法及其工具,就可利用所有保信系统内都具有的文件 传输协议来直接传输,在主站内就可免去解压,直接进 行存储,需要分析的时候,只需调用相应的解压软 件。这种应用模式可把以往算法和传输的紧密关系 完全解耦,更加简单可靠,而各种第三方分析软件系

收稿日期:2011-11-26:修回日期:2013-03-15

基金项目:中央高校基本科研业务费专项资金资助项目(ZYGX-2009J089)

Project supported by the Fundamental Research Funds for the Central University(ZYGX2009J089)

统也可携带此工具,为各种电力用户节约存储空间。 基于这种思路,笔者在已有研究的基础上,结合多 年的电力系统故障分析软件研发经验,提出了面向 COMTRADE标准的并行压缩/解压算法,并构建了 相应的压缩/解压工具。算法根据 COMTRADE中不 同种类记录信息的特点,采用不同的压缩算法或混合 压缩算法,获得极大的压缩比,且具有线性加速比, 可较好地解决以往算法在实际应用中出现的问题。

## 1 IEEE COMTRADE 格式标准

IEEE Std C37.111—1999 规定和记录信息相关 的文件有 4 个:头标文件(文件扩展名为 HDR)、配 置文件(文件扩展名为 CFG)、数据文件(文件扩展名 为 DAT)以及 1 个可选的信息文件(文件扩展名为 INF)。其中对于第三方暂态数据分析软件而言最重 要的是配置文件和数据文件,其他文件是可选的。对 于压缩算法而言,除了数据文件外,其他文件都可视 为纯文本文件,以文本流的方式进行压缩。

#### 1.1 配置文件

要求计算机程序能够读取配置文件,获取内部 的属性信息,并采用这些属性信息正确地读取数据文 件(\*.DAT)中的记录数据。配置文件为ASCII 文本 文件,包含着计算机程序为了正确解读数据文件而需 要的信息。配置文件由若干行组成,每个字段都有 对应的明确物理意义。其具体内容为:站名,记录装 置的特征,COMTRADE标准的修改年份;模拟量和状 态量通道的数量;模拟量通道采样信息(包括模拟通 道索引号、通道识别名称等配置信息);状态量通道 采样信息(包括状态通道索引号、通道识别名称等配 置信息);被采样线路的频率;采样频率个数,实际采 样频率及其对应的结束采样下标;第1个数据点的日 期和时间;触发点的日期和时间;数据文件类型,表 示数据文件是ASCII 文本文件还是二进制文件;时间 标记倍乘系数。

#### 1.2 数据文件

数据文件记录着每个采样通道中的每个采样数 值。数据文件可以是 ASCII 或二进制格式。二进制 数据文件和 ASCII 数据文件格式类似,每行应分为 *T*+2列,其中*T*是配置文件中模拟量通道和状态量 通道的总和,另外2列是采样序号和时间标记。二 进制文件以 BIT 形式集中存放状态量通道数据。各 个数据之间没有分隔符,每组采样值之间没有回车 换行符隔开。各列的具体内容如下:第1列为采样 序号;第2列为采样数据的时间标记;第3组的列为 表示模拟量通道信息的采样数据值;第4组的列为 表示状态量通道信息的采样数据值。其数据文件格 式的例子如图1所示。

1,-100000,-8025,6067,6059,2022,1,0,1 2,-999900,-7925,6167,6159,2234,1,0,1				
采样 相对设备 序号 记录起始点 的时间(μs)	模拟量通道1 到4的采样值	状态量通道1 到3的状态值		
图 1	数据文件格式	:		

#### Fig.1 Format of data file

# 2 IEEE COMTRADE 数据文件压缩/解压 算法

从图 1 可见数据文件分为采样序号、时间标记数据、模拟量通道数据、状态量通道数据 4 个部分。其中采样序号在压缩时可以完全丢弃,下面分别讨论其他 3 个部分的压缩算法。

## 2.1 时间标记数据压缩算法

由于采样时间标记是一个递增的量,可通过配置 文件中的采样频率间接计算得到,因此压缩此项数据 时只需记录数据文件中记录的第1个时间标记数 据、配置文件中的采样率个数和实际采样频率及其对 应的结束采样下标、配置文件中的时间标记乘数3种 信息,即可在解压时进行无损恢复。解压算法公式为:

$$t_{s} = \operatorname{mod}[t_{1} + (t_{a}/t_{m})]$$

$$t_{a} = 10^{6} n/f$$
(1)

其中,t<sub>a</sub>为累加偏移时间,初值为0;f为当前采样频 率;n为当前采样频率下从0开始的序号;t<sub>1</sub>为数据 文件中的第1个采样时间;t<sub>m</sub>为配置文件中的时间 标记乘数;t<sub>s</sub>为采样时间标记。算法首先计算累加偏 移时间,并将采样间隔时间单位转换为 COMTRADE 标准规定的微秒。将累加偏移时间除以时间标记乘 数后加上t<sub>1</sub>,再取整就可无损恢复出所有的时间标记。

## 2.2 状态量数据压缩算法

在 COMTRADE 标准中, BIANRY 格式数据文件 中的状态量数据是以 BIT 形式保存的, 而在 ASCII 码 中则是以 1 个字符("0"或"1")来表示。在 BINARY 格式中, 状态量通道所占存储空间的计算公式为:

$$S_{\rm D} = \begin{cases} 2[\mod(C/16) + 1] & C\%16 \neq 0\\ 2\mod(C/16) & C\%16 = 0 \end{cases}$$
(2)

其中,mod为取整运算,%为取余数运算。

由式(2)可以知道,即使是只有1个状态量通 道,状态量数据也要占据2Byte的存储空间,即对于 COMTRADE标准而言状态量数据保存是有冗余量 的。并且通过电力系统内大量COMTRADE标准历 史数据的验证,在所有状态量通道中,95%以上的通 道记录根本没有状态变化,而只有极少量的状态量通 道记录了开关量变位信息。

对于这种只有 2 种变化的量,结合 COMTRADE 状态量的特点,算法设计了最简游程编码 RLE(Run Length Encoding),即采用游程 L 的正表示状态 1,L

的负表示状态 0,而 L 的绝对值表示游程长度,例如 对于 111111110000000,就可用(+8,-7)来表示,而 对于没有状态量变换的状态量通道,将其分别归为全 1 状态量通道和全 0 状态量通道类,其中每一类只是 记录对应的通道编号,通道编号的整型字节长度依据 通道数量的多少而定,通常情况下数字量通道数都在 256 以内,所以可以将通道编号整型长度的初值设为 1 Byte。通过这种压缩编码就可对具有大量状态量 信息的 COMTRADE 标准记录文件进行大幅压缩。

#### 2.3 高频采样模拟量数据压缩算法

高频采样模拟量数据在 IEEE COMTRADE 数据文件中是以 2 Byte 的整型数保存的。因此,基于浮点数运算的 Mallat 算法将无法实现无损压缩,且计算量也比第 2 代提升格式的离散小波变换至少高出一倍。因此本算法中针对模拟量数据的压缩首先采用提升格式的离散小波变换,根据用户选择是否进行有损压缩,进一步采用硬阈值方法进行量化<sup>[7]</sup>,最后采用熵编码算法进行压缩的混合压缩算法。

2.3.1 提升格式小波变换

提升格式小波变换与 Mallat 算法相比的区别在 于不依赖傅里叶变换,在时域内就可实现小波构造。 对信号进行一次提升变换就相当于进行一次小波分 解,一个典型的提升格式离散小波变换包括分裂、预 测和更新 3 个步骤。而重构则是此过程的逆,即更 新、预测和合并。整个提升小波的分解和重构如图 2 所示。





从图 2 可知,提升格式小波变换就是不断对偶数序列(把原始序列也看作偶数序列)减半后进行分解的过程,而重构过程则是分解过程的逆。对应于 计算内部实现,对其分解和重构的原理如图 3 所示。



从图 3 可知,小波分解的第 1 步分裂,就是将原

始序列 S<sub>i</sub>的偶数部分移动到前半部分,奇数部分移 动到后半部分,而后进行预测和更新,一直持续该过 程,而重构是此过程的逆,不需要任何格外的内存就 可完成全部变换。对于一个有 2<sup>H</sup> 个点的输入原始 数据,经过 h(h<H)次小波变换后,尺度系数的点为 序列左边的 2<sup>H-h</sup> 个点,其余点为小波系数,由此可见 小波分解层数越深,可供量化的小波系数也就越 多,故压缩效率也就更高。本文算法中提供了多种 整数小波变换,可供用户进行选择,它们在尺度 j 上 的变换公式如下所示。

CDF(2,2)插值小波变换:

$$D_{j,l} = D_{j,l} - \text{mod}[(S_{j,l} + S_{j,l+1})/2 + 0.5]$$

$$S_{j,l} = S_{j,l} + \text{mod}[(D_{j,l-1} + D_{j,l})/4 + 0.5]$$

$$CDF(4,2) \\ \pm \text{fid} \wedge \& \& \& \& \& B_{j,l} = D_{j,l} - \text{mod}[9(S_{j,l} + S_{j,l+1})/16 - (S_{j,l-1} + S_{j,l+2})/16 + 0.5]$$

$$CDF(6,2) \\ \pm \text{fid} \wedge \& \& \& \& \& \& \& \& \& \& \& B_{j,l} = S_{j,l} + \text{mod}[(D_{j,l-1} + D_{j,l})/4 + 0.5]$$

$$CDF(6,2) \\ \pm \text{fid} \wedge \& \& \& \& \& \& \& \& \& \& \& \& \& \& \& \& \& B_{j,l-1} - \text{mod}[75(S_{j,l} + S_{j,l+1})/128 - 25(S_{j,l-1} + S_{j,l+2})/256 + 3(S_{j,l-2} + S_{j,l+3})/256 + 0.5]$$

$$(5)$$

$$D_{j,l} = D_{j,l} - \text{mod} [225(S_{j,l} + S_{j,l+1}) / 2048 - 245(S_{j,l-1} + S_{j,l+2}) / 2048 + 49(S_{j,l-2} + S_{j,l+3}) / 2048 - (6)$$
  

$$5(S_{j,l-3} + S_{j,l+4}) / 2048 + 0.5]$$

 $S_{j,l} = S_{j,l} + \text{mod}[(D_{j,l-1} + D_{j,l})/4 + 0.5]$ 

其中,S<sub>j</sub>、D<sub>j</sub>为经过 j-1 层分解后得到的偶序列和奇 序列,l为序列数组下标。

2.3.2 熵编码压缩算法

 $D_{i,l}$ 

熵编码压缩主要是利用数据或数据序列出现概率的分布特性来寻求概率与码字长度间的最优匹配。常用的熵编码有Huffman、LZ77、LZ78、LZW、RLE、算术编码等<sup>[17]</sup>。众多文献已经证明了单纯采用一种熵编码很难取得高压缩比,文献[12]采用了RLE和LZW的混合编码,取得了较好的压缩效果。本文中采用了著名的混合熵编码Deflate算法,其采用了经过优化后的LZ77和Huffman的混合熵编码算法,压缩比和速度更快。由于Deflate算法很好地平衡了压缩效率和速度,所以很多压缩软件(PKZIP、WinZip、gzip)均采用了Deflate算法或由其衍生出来的其他算法,Deflate算法的解压算法名称为Inflate。

LZ77 编码算法是由 Lempel 和 Ziv 于 1977 年提 出的一种字典压缩算法,其核心思想是把已输入的数 据流的一部分作为字典,编码器为输入流开一个滑动 窗口,随着字符串编码的输入把窗口中的数据从右向 左移动,并在其中寻找数据中相同的部分,即寻找最 长字符串匹配,由此在编码过程中形成一种内容形式 为(距离,匹配长度)的动态字典。LZ77 在压缩时 需要进行大量的字符串匹配工作,这也是 LZ77 效率 上最低的地方,为此 Deflate 算法采用哈希表来提高 匹配速度。在哈希表匹配过程中,采用当前字符串 中头 3 个字符来计算 1 个哈希值,并将具有相同哈希 值的匹配字符串利用链表连接起来,而寻找最长匹配 字符串的过程,即通过哈希值找到链表头后,遍历链 表,最后确认最长匹配字符串。

在经过 LZ77 编码输出后,Deflate 算法采用 Huffman编码进一步压缩数据。Deflate 算法在对一 块数据进行 Huffman编码前,会同时建立静态 Huffman 树和动态 Huffman树;然后根据要输出的内 容和生成的 Huffman树,计算静态和动态 Huffman树 的最终生成块大小,而后进行比较,采用生成块较小 的算法进行编码。采用这种方法,Deflate 算法避免 了由于文件较小时动态 Huffman树编码比静态 Huffman编码生成块大的情况,保证最终输出块始终 小于编码输入块,这一点是其他算法很难做到的。

2.4 低频采样模拟量数据压缩算法

在电力系统内部,很多 COMTRADE 标准记录文件中的最后一个采样段都是低频采样的(小于 100 Hz) 模拟量数据,这些低频数据通常是当前模拟量通道 的有效值。这些数据对于故障分析没有太大意义, 并且通常也是没有任何变化的量,对于这类低频数据 本算法缺省采用 RLE 算法,将数据压缩成(数值,长 度)对。

## 3 最佳小波分解尺度下的最佳数据区间划 分算法

由于 COMTRADE 标准记录文件内通常包含多 个采样频率下的数据,压缩算法不可能一次性加载全 部模拟量通道数据,并且待压缩数据点数也不太可能 是 2 的幂级数,因此可能需要在采样频段尾部补零, 压缩算法就面临一次读取多少采样数据进行小波分 解、压缩的问题。为此本文提出了一种采用配置文 件中提供的采样频率信息的压缩数据区间最佳划分 算法。

为防止小波分解层数过多而导致基频分量在小 波变换过程中被分解,影响压缩效率,本文算法采用 式(7)<sup>171</sup>计算最佳小波分解层数,并由此倒推出要获 取最佳压缩效果的最小数据点数公式见式(8):

$$n = \mod\left(\log_2 \frac{f_s}{50} - 1\right) \tag{7}$$

$$B=2^{n+1} \tag{8}$$

式(8)的意义是在采样频率 f<sub>s</sub>下,经过 n 层小波 分解,最后刚好有 2 个点为尺度系数,即实际中要 求的待分析数据量必须为 B 的 2<sup>k</sup>(k=0,1,2,…)倍, 才能获取最佳压缩效果。根据式(8)可得到在采样 频率为 f<sub>s</sub>、采样数据总数为 K 时的最小补零个数为:

$$n_0 = B - K \% B \tag{9}$$

假设处理数据缓冲区大小为 $S_B=2^m(m>n+1)$ ,于是可得以 $S_B$ 划分的数据区间个数M为:

$$M = \operatorname{mod}\left[ \left( K + n_{0} \right) / S_{\mathrm{B}} \right]$$

$$(10)$$

而对余下的数据则需要计算出其相对于 B 的 倍数 L。

$$L = [(K + n_{o}) \% S_{B}] / B$$
(11)

对式(11)中的L以二进制方式从最高 BIT 位起 依次判断P位是0还是1,如果是1则利用式(12)获 取一个数据划分大小Y:

$$Y = 2^{P}B \tag{12}$$

通过以上算法就可得到一组可最少补零的数据 划分,例如在f<sub>s</sub>=9600 Hz时,共采集了15543 点数 据,当前计算缓冲区为4096个整型数据,则采用以 上算法可自动将数据划分为3个大小为4096的区 间、1个大小为2048的区间、1个大小为1024的区 间和1个大小为256的区间。最小的一个区间需要 补零73个。由此可见采用此算法所划分的区间,满 足了最佳小波分解层数的要求,且补零个数最少,在 压缩效果和压缩效率间找到了一个较好的平衡点。

### 4 并行压缩算法

如图 4 所示,并行压缩算法首先解析 COMTRADE 配置文件,再利用其内信息创建压缩文件头并写入压 缩文件后,将 COMTRADE 配套文本文件(\*.cfg等) 依次采用 Deflate 算法进行压缩,并写入压缩文件中。 随后通过第 3 节提供的划分算法,依据采样率信息自 动创建数据压缩划分区间,并根据此区间依次将原始 数据读入并解析到模拟量通道数据压缩共享管理器 内预先分配好的通道数据缓冲区中。此时通知激活 已经创建并等待的模拟量压缩线程组和状态量压缩 线程组(并行压缩算法采用的线程 CPU 核心映射原 则是:每个核心映射 1 个模拟量压缩线程和 1 个状态 量压缩线程)。而后这些线程将向模拟量通道数据 压缩管理器请求待压缩通道序号。

对于模拟量数据而言,压缩线程通过判断采样频 率是否高于 100 Hz 选用不同的压缩策略。对于有损 压缩,当采用提升格式小波变换后,采用文献[7]中 介绍的阈值处理方法,该方法在阈值选择过程中考虑 了噪声的小波变换系数在小波系数空间的传播特 性,能够很好地抑制噪声对压缩性能的影响。压缩后 的数据直接写入压缩文件内。

$$t_i = \gamma \sigma \sqrt{2 \log N} \tag{13}$$

其中, $t_i$ 为在尺度j下的硬阈值; $\sigma$ 为噪声强度;N为



图 4 COMTRADE 并行压缩算法

Fig.4 COMTRADE parallel compression algorithm

尺度j下的小波系数个数; $\gamma$ 为常数,选择范围为 1~3。 其中噪声强度 $\sigma$ 可根据第i次小波变换后的小波 系数计算得到,计算公式为:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (d_i - d_i)^2}$$
(14)

其中,*d<sub>i</sub>*为第*i*次小波变换后的小波系数,*d*为均值, N为小波系数个数。本算法中,γ取值1、2、3分别被 映射为"一般"、"较好"、"最好"3种压缩选项。

对于状态量数据而言,压缩线程每次更新在状态 量通道数据缓冲区的压缩数据,只有当压缩线程遍历 状态量通道的全部状态量数据后,才提炼得出全1通 道序号数组、全0通道序号数组和状态量发生变化通 道压缩数据,最后将这3类压缩数据写入文件。

这种基于申请序号的线程竞争计算方式,可使压 缩算法效率随着 CPU 核心数量的增加而提高。

## 5 压缩文件格式

经过第4节的并行压缩算法,最终压缩文件形成 了如图5所示的压缩存储格式,其中每个部分都是一 个可根据压缩数据变化的量,不能直接采用固定大小 结构体设计,为此设计了如下所示的通用变长结构体。 struct AInfo

\_int 32 nStructSize;

\_int 16 nStructType;

//其他信息

type Data[1];//某类型数组声明

此结构的大小计算程序公式为:

nStructSize = sizeof(AInfo) +

sizeof(type)\*(Count-1) (15)
其中,Count为实际存储数据大小,采用此大小分配内
存后,对于 type 类型数据 Data,可直接用数组形式
访问,这样就形成了一个变长的结构体。

	压缩文件头
	文件名
	COMTRADE 配套 TEXT 文件1 压缩数据
I	:
	COMTRADE 配套 TEXT 文件 S 压缩数据
1	模拟量通道1压缩数据
第1组	模拟量通道2压缩数据
数据划分	E
↓ ↓	模拟量通道 F 压缩数据
	÷
1	模拟量通道1压缩数据
第 R 组	模拟量通道2压缩数据
数据划分	:
Ļ	模拟量通道 G 压缩数据
1	全1状态量通道编号
山大旦	全0状态量通道编号
1八心里 压缩数据	有变化状态通道压缩数据
	:
Ļ	有变化状态通道压缩数据

图 5 压缩文件格式 Fig.5 Format of compressed file

压缩文件的读取就是利用 nStructSize 来确定读

写大小,采用 nStructType 来判断数据类型,进而准确读取所有的压缩数据块。在此压缩存储格式中, 模拟量数据以压缩时数据划分的形式顺序保存,而 状态量数据则集中保存在文件最后,在压缩文件头中 有指向状态量压缩数据的具体偏移数据。

#### 6 并行解压算法

对于第5节阐述的压缩文件格式,采用如图6所示的并行解压算法进行解压。在解压主线程内,首 先采用 Inflate 算法对 COMTRADE 配套文本文件压 缩数据进行解压,并保存到文件。之后解压主线程 通过压缩文件头内提供的状态量压缩数据偏移量信 息,直接定位并一次性将压缩数据读入到状态量压缩 数据共享存储器内。随后依据压缩文件头内提供的 数据划分信息,将一次数据划分的全部数据读入模拟





#### 图 6 并行解压算法

Fig6 Parallel decompression algorithm

量通道压缩数据共享管理器内,之后解压主线程将激 活模拟量通道数据解压线程,并等待解压完毕(并行 压缩算法采用的线程 CPU 核心映射原则是每个核 心映射1个模拟量解压线程)。激活后的各个解压 线程向共享管理器申请解压通道序号,根据原始采样 率判断是否为高频模拟量压缩数据,之后采用不同的 解压策略。当一次数据划分并行解压完毕后,解压 主线程等待结束,随后采用状态量压缩数据解压出当 前数据划分下的状态量信息,再将时间信息还原后加 上采样编号,最后写入解压文件。持续此过程直到 所有数据划分区间被读取并被并行解压完毕。同第 4节所述并行压缩算法类似,此并行解压算法采用基 于申请序号的线程竞争处理方式,可使解压效率随着 CPU 核心数量的增加而提高。

#### 7 试验

将本文算法应用于4核CPU计算机,配置为:单核主频2.4GHz,内存1GByte,硬盘转速7200r/min。

从表 1 可见 Deflate 算法的速度很快,而其解压 算法 Inflate 的速度是 Deflate 算法是 5~9 倍。从表 2 可见,对于同样大小的原始计算数据,不同插值点 的提升格式小波变换计算时间的差别并不明显,以

表 1	Deflate / Inflate 算法效率试验结果	
Tab.1	Results of Deflate/Inflate algorithm	ı
	efficiency test	

enciency test				
压缩块大小/KByte	压缩时间/μs	解压时间/μs		
8	310	57		
16	420	75		
32	870	110		
64	1450	160		

#### 表 2 提升格式小波变换效率试验结果

Tab.2 Results of lifting wavelet transform efficiency test

小油米刑		计算时	†间/μs	
小伋天室	8×10 <sup>3</sup> 点	1.6×10 <sup>4</sup> 点	3.2×10 <sup>4</sup> 点	6.4×10 <sup>4</sup> 点
CDF(2,2)	82	160	371	737
CDF(4,2)	84	165	390	745
CDF(6,2)	97	198	443	870
CDF(8,2)	105	210	463	945

6.4×10<sup>4</sup> 个计算点为例,CDF(2,2)和 CDF(8,2)计算时间相差 208 ms。

为了体现算法在实际应用中的压缩效果,选用了 在电力系统内部具有典型意义的4种 COMTRADE 文件(文件详细信息如表3所示),其具体分类如下 所表述。

文件1:典型的单相接地故障记录数据文件,记录末尾有少量低频采样数据。

文件 2:典型的保信系统上传文件,记录大量意 义不大的稳态数据及其开关量变位信息,记录末尾 有较多低频采样数据。

文件 3:故障持续时间较长,采样频率高,记录 中没有低频数据。

文件4:系统发生较长时间扰动,录波器连续进 行长时间录波,完整记录整个过程,此类连续长时间 采样也是未来电力系统自动化录波的发展趋势。

为和以往文献试验结果对比,试验采用以往众多 文献推荐的 CDF(4,2)小波,从表4、5 所示的压缩比

表 3 压缩/解压试验 COMTRADE 数据 Tab.3 COMTRADE data of compression/ decompression test

文件	文件大小/KByte			模拟量	壮态量	通道采样
编号	BINARY 数据	ASCII 数据	配套 文本	通道数	通道数	点个数
1	57	133	0.7	8	7	2 210
2	248	616	1.1	13	16	7 082
3	936	2 935	2.2	18	32	19 973
4	15 652	46 490	3.9	33	64	195 465

表 4 有损压缩试验结果

	Tab.4 Results of loss compression test				
1	文件	压缩文件	相对于 BINARY	相对于 ASCII	
	编号	大小/KByte	原始文件的压缩比	原始文件的压缩比	
	1	4.2	13.6	31.7	
	2	5.8	42.8	106.2	
	3	16.4	57.1	179.0	
	4	162.0	96.6	287.0	

表 5 无损压缩试验结果

	Tab.5 Resul	ts of lossless comp	pression test
文件	压缩文件	相对于 BINARY	相对于 ASCII
编号	大小/KByte	原始文件的压缩比	原始文件的压缩比
1	18.4	3.1	7.2
2	33.5	7.4	18.4
3	246.0	3.8	11.9
4	2 745.0	5.7	16.9

试验结果上看,本文算法均可获得非常大的压缩比, 当 COMTRADE 文件逐渐变大后,压缩比也逐渐变 大,特别是文件 4,在有损压缩时取得了高达 96.6 和 287.0 的压缩比。而对于无损压缩,此算法也取得了 很好的效果,文件 4 取得了 5.7 和 16.9 的高无损压缩 比,而文件 2 由于尾部有较多低频数据,取得了更高 的压缩比。

表 6 中针对文件 4 类型的 COMTRADE 文件, 采用 4 种小波分别进行有损压缩,可见随着参与插值 的点数增加,获得了更小的压缩文件。笔者还通过大 量试验证明,在诸多文献中提到的 CDF(4,2)的压缩 效果最好的结论,实际上过于片面,试验中 CDF(6,2)、 CDF(8,2)小波对于 COMTRADE 中大量存在的小扰 动系统录波数据具有更高的压缩比。并且由表 2 可 见插值点数的增加对于提升小波变换速度的影响有 限,所以笔者认为对于实际中的 COMTRADE 文件而 言,推荐采用 CDF(6,2)和 CDF(8,2)这类插值点更 多的小波,使得变换后得到的小波系数更加平滑,有 利于压缩。

表 6 不同小波选择的压缩结果

Tab.6	Results	of	$\operatorname{compression}$	$\mathbf{b}\mathbf{y}$	different	wavelet
-------	---------	----	------------------------------	------------------------	-----------	---------

小油米到	有损压缩	无损压缩
小放矢型	文件大小/KByte	文件大小/KByte
CDF(2,2)	3 572	251
CDF(4,2)	3 501	162
CDF(6,2)	3 479	152
CDF(8,2)	3 468	148

表 7 显示了对 4 种不同文件的内故障记录通道 有损压缩后和原始通道数据计算均方误差,采用的 小波为 CDF(8,2)。从表中可见有损压缩后的解压 数据和原始数据误差都在小数点后 3 位,满足有损 压缩后的可用要求。

#### 表 7 故障通道数据有损压缩最大均方误差

Tab.7 Maximum deviation of loss compression for different faulty channels

		2	
立他编早	Ē	是大均方误差/9	%
又口畑ゥ	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
1	0.21	0.26	0.29
2	0.23	0.25	0.31
3	0.31	0.36	0.38
4	0.44	0.46	0.49

表 8 中采用文件 4 类型 COMTRADE 数据在 2 核、3 核、4 核计算机上进行了加速比试验。可见,此

表 8 并行压缩/解压加速比试验结果
Tab.8 Results of parallel compression/
decompression acceleration ratio test

decompression acceleration ratio test				
运行	串行	执行时间/ms,加速比		
类型	时间/ms	2 核	3 核	4 核
压缩	1 562	849,1.84	592,2.64	482,3.24
解压	754	412,1.83	289,2.61	234,3.22

算法可随着未来 COMTRADE 数据文件的加大、 CPU核心数的增加,获取到线性加速比。

#### 8 结语

以往压缩算法文献都只是针对模拟量数据进行 讨论的,而完全忽视了电力系统内绝大部分录波数据 都是以 IEEE COMTRADE 格式进行保存的情况,并 且由于电力系统通信网的快速发展,原有算法的应用 背景也发生较大变化,原有研究已经不能适应电力系 统的发展,因此本文提出了直接面向 IEEE COMTRADE 格式的并行压缩/解压算法,并构建了相应的通用压 缩/解压工具,大量试验证明对于绝大部分现场运行, COMTRADE 文件都可获得很大的压缩比,并且算法 适应了多核并行计算的发展趋势,具备线性加速比。

#### 参考文献:

- [1] 白青刚,夏瑞华,周海斌,等.采用高性能集成芯片的故障录波故障设计[J].电力系统自动化,2005,29(22):94-96.
  BAI Qinggang,XIA Ruihua,ZHOU Haibin, et al. Design of fault wave recording device using high performance integrated microchip[J]. Automation of Electric Power Systems,2005,29(22): 94-96.
- [2] SANTOSO S, POWERS E J, GRADY W M. Power quality disturbance data compression using wavelet transform methods [J]. IEEE Trans on Power Delivery, 1997, 12(3):1250-1256.
- [3] LITTLER T B, MORROW D J. Wavelets for the analysis and compression of power system disturbances[J]. IEEE Trans on Power Delivery, 1999, 14(2):358-364.
- [4]何正友,钱清泉,刘志刚. 一种基于优化小波基的电力系统故障 暂态数据压缩方法[J]. 中国电机工程学报,2002,22(6):1-5.
   HE Zhengyou,QIAN Qingquan,LIU Zhigang. An electric power fault transient data compression method based on optimal wavelet[J]. Proceedings of the CSEE,2002,22(6):1-5.
- [5] 刘志刚,钱清泉.基于多小波的电力系统故障暂态数据压缩算法[J].中国电机工程学报,2003,23(10):22-26.
  LIU Zhigang,QIAN Qingquan. Compression of transient data in electric power system based on multi-wavelet[J]. Proceedings of the CSEE,2003,23(10):22-26.
- [6] 胡国胜,任震,黄雯莹,等. 基于最小带宽离散小波的故障信号 消噪与压缩[J]. 中国电机工程学报,2001,21(11):6-8.
  HU Guosheng,REN Zhen,HUANG Wenying, et al. Fault signals denosing and compressing based on MBDW filters[J]. Proceedings of the CSEE,2001,21(11):6-8.
- [7] 乐全明,郁惟镛,柏传军,等. 基于提升算法的电力系统故障录波数据压缩新方案[J]. 电力系统自动化,2005,29(5):74-78.

YUE Quanning, YU Weiyong, BAI Chuanjun, et al. Novel compression scheme of fault recording data in power systems based on lifting algorithm[J]. Automation of Electric Power Systems, 2005, 29(5):74-78.

- [8] 苗世洪,王少荣,刘沛,等.数据压缩技术在电力系统通信中的应用[J].电力自动化设备,1999,19(3):32-33.
  MIAO Shihong,WANG Shaorong,LIU Pei,et al. Research on application of data compression technology in electric power system communication[J]. Electric Power Automation Equipment, 1999,19(3):32-33.
- [9] 张超,房若季.改进的 LZSS 压缩算法在故障信息文件远传中的应用[J]. 电网技术,2003,27(6):42-44.

ZHANG Chao, FANG Ruoji. Application of improved LZSS algorithm in remote transmission of fault information files [J]. Power System Technology, 2003, 27(6):42-44.

- [10] 闫常友,杨奇逊,刘万顺. 基于提升格式的实时数据压缩和重构 算法[J]. 中国电机工程学报,2005,25(9):6-10.
  YAN Changyou,YANG Qixun,LIU Wanshun. A real-time data compression & reconstruction method based on lifting scheme
  [J]. Proceedings of the CSEE,2005,25(9):6-10.
- [11] 任明明,张彦斌,贾立新. 一种基于提升小波和零树编码的录波 数据压缩算法[J]. 西安交通大学学报,2006,40(4):494-496.
   REN Mingming,ZHANG Yanbin,JIA Lixin. Compression method based on lifting wavelet transform and embedded zero-tree coding[J]. Journal of Xi'an Jiaotong University,2006,40(4):

494-496.

- [12] 周瑞,鲍文,于霄,等. 基于提升小波和混合熵编码的数据压缩 方法[J]. 电力系统自动化,2007,31(22):65-79.
  ZHOU Rui,BAO Wen,YU Xiao,et al. Data compression method based on lifting wavelet transform and hybrid entropy coding [J]. Automation of Electric Power Systems,2007,31(22):65-79.
- [13] HSIEH C T, HUANG S J. Disturbance data compression of a power system using the Huffman coding approach with wavelet transform enhancement[J]. IEEE Proc-Gener Transm Distrib, 2003,150(1):7-14.
- [14] MALLAT S. A theory for multiresolution signal decomposion: the wavelet representation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1989, 11(7):674-693.
- [15] SWELDENS W. The lifting scheme:a construction of secondgeneration wavelets[J]. SIAM Journal of Mathematical Analysis, 1998,29(2):511-546.
- [16] IEEE Standards Board. IEEE C37.111-1999 COMmon format for TRAnsient Data Exchange (COMTRADE) for power systems
   [S]. [S.I.]:Power System Relaying Committee of the IEEE Power Engineering Society, 1999.
- [17] 吴乐南. 数据压缩[M]. 北京:电子工业出版社,2000:43-66.

#### 作者简介:

桂 勋(1978-),男,贵州安顺人,副教授,博士,研究方 向为电力暂态信息处理(E-mail:guinh3@163.com)。

# Parallel compression and decompression algorithm for massive recording data in IEEE COMTRADE format

GUI Xun

(School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 610054, China)

**Abstract**: A parallel compression/decompression algorithm is put forward for the massive fault recording data in IEEE COMTRADE format. A lossless data recovery formula is given for the time information in COMTRADE data file and an optimized coding algorithm is proposed for the status data. The high-frequency analog data are transformed with the lifting wavelet, quantified with hard-threshold, and then compressed by the Deflate algorithm based on the LZ77-Huffman hybrid entropy coding while the low-frequency analog data by the RLE coding algorithm. A data partitioning algorithm is proposed to effectively compress the COMTRADE data file in turn, which is based on the optimal layers of wavelet decomposition for high-frequency analog data for patching the least zeros. Mapped to different CPU cores, the compression/ decompression threads access the computing data through the competition of channel serial number. The format of compressed file and the technology of variable-length storage and access are introduced. Test indicates that, the great compression ratio is gained and the linear acceleration ratio is obtained along with the increase of CPU core quantity and COMTRADE file size.

**Key words**: data recording; IEEE COMTRADE; compression; decompression; compression; decompression; data processing; LZ77; Huffman; Deflate; wavelet transforms

143