SIMD-Based Decoding of Posting Lists

Alexander A. Stepanov, Anil R. Gangolli, Daniel E. Rose
Ryan J. Ernst, Paramjit S. Oberoi

A9.com
130 Lytton Ave.
Palo Alto, CA 94303

{stepanov,gangolli,danrose,rjernst,paramjit}@a9.com

ABSTRACT

Powerful SIMD instructions in modern processors offer an opportunity for greater search performance. In this paper, we apply these instructions to decoding search engine posting lists. We start by exploring variable-length integer encoding formats used to represent postings. We define two properties, *byte-oriented* and *byte-preserving*, that characterize many formats of interest. Based on their common structure, we define a taxonomy that classifies encodings along three dimensions, representing the way in which data bits are stored and additional bits are used to describe the data. Using this taxonomy, we discover new encoding formats, some of which are particularly amenable to SIMD-based decoding. We present generic SIMD algorithms for decoding these formats. We also extend these algorithms to the most common traditional encoding format. Our experiments demonstrate that SIMD-based decoding algorithms are up to 3 times faster than non-SIMD algorithms.

Categories and Subject Descriptors

E.4 [**Coding and Information Theory**]: Data Compaction and Compression; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods*; C.1.2 [**Processor Architectures**]: [Single-instruction-stream, multiple-data-stream processors (SIMD)]

General Terms

Algorithms, Performance, Measurement, Experimentation

Keywords

variable-length integer encoding, SIMD

1. INTRODUCTION

The central data structure in search engines is the *inverted index*, a mapping from index terms to the documents that contain them. The set of documents containing a given word is known as a *posting list*. Documents in posting lists are typically represented by unique nonnegative integer identifiers. Posting lists are generally kept in sorted order to enable fast set operations which are required for query processing.

Because posting lists typically account for a large fraction of storage used by the search engine, it is desirable to compress the lists. Smaller lists mean less memory usage, and in the case of disk-based indices, smaller lists reduce I/O and therefore provide faster access. A common way to compress posting lists is to replace document IDs in the list with differences between successive document IDs, known as Δ-*gaps* (sometimes just *gaps* or *d-gaps*). Since Δ-gaps are on average necessarily smaller than raw document IDs, we can use a variable-length unsigned integer encoding method in which smaller values occupy less space. In a sorted list the Δ-gaps are always non-negative, so we are only concerned with encoding nonnegative integers; for the remainder of the paper "integer" means unsigned integer.

The integer encoding for a posting list is performed infrequently, at indexing time. The decoding, however, must be performed for every uncached query. For this reason, efficient integer decoding algorithms are essential, as are encoding formats that support such efficient decoding.

We started by exploring several commonly-used variable-length integer encodings and their associated decoding algorithms. Based on their common structure, we defined a taxonomy that encompassed these existing encodings and suggested some novel ones. Our investigation was motivated by the desire to incorporate fine-grained parallelism to speed up integer decoding. We were able to develop decoding methods that use SIMD instructions available on many general-purpose processors, in particular current Intel and forthcoming AMD processors. In this paper, we present these methods and evaluate them against traditional techniques. Our results indicate significant performance benefit from applying SIMD to these new formats.

The remainder of the paper is organized as follows. We review some related work on encoding formats and decoding algorithms in Section 2. Section 3 presents our taxonomy of byte-oriented encodings. Section 4 describes three particular encoding formats which are well suited to SIMD parallelism in more detail. Section 5 gives an overview of our use of SIMD on Intel-compatible processors. In Section 6 we provide SIMD algorithms for decoding the formats introduced in Section 4. Section 7 contains an evaluation of the results, followed by our conclusions in Section 8.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

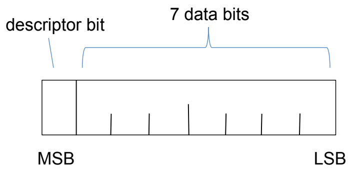

Figure 1: Format known in the literature as "vbyte", "vint", etc. Called *varint-SU* in the taxonomy of Section 3.

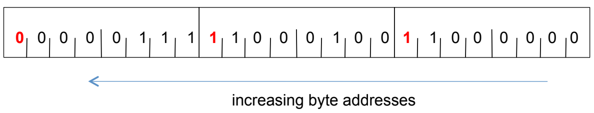

Figure 2: Storing the integer 123456 in varint-SU format.

2. RELATED WORK

The problem of integer encoding has been studied for several decades. General compression techniques such as Huffman coding often utilize analysis of the data to choose optimal representations for each codeword. Many applications, however, use techniques that do not depend on data distribution. These are called *nonparametric* codes [4].

Encoded representations of data may also be classified as byte- or bit-aligned, depending on whether codewords are required to end on byte boundaries. In addition, Anh [1, 2] introduced an interesting class of *word-aligned* encodings (e.g. Simple-9). These encodings occupy an intermediate place between bit-aligned and byte-aligned encodings by allowing codewords to end at an arbitrary bit position as long as they do not cross the machine word boundary.

While there are multiple choices for encoding posting lists in information retrieval applications, in this paper we concentrate exclusively on non-parametric, byte-oriented encodings.

The most common such encoding format uses 7 bits of each byte to represent data, and one bit to represent whether the encoded representation of the integer continues into the next byte (see Figure 1). This format has a long history, dating back at least to the MIDI specification in the early 1980s [14], and foreshadowed by earlier work on byte-based vocabulary compression by H.S. Heaps [11]. As early as 1990, Cutting and Pedersen [7] used this format to encode Δ-gaps in an inverted index. The information retrieval literature refers to the format by many different names, including vbyte, vint, and VB. Different authors give slightly different versions that vary on endianness, location of the continuation bit (most significant vs. least significant), and whether 0 or 1 indicates continuation. Figure 2 illustrates this encoding for the integer 123456 with little-endian, most significant bit and 1 for continuation.

Another format *BA*, introduced by Grossman in 1995 [10] uses 2 bits of the first byte to indicate the length of the encoded integer in binary. In his 2009 WSDM keynote, Dean [8] described a format he calls *group varint* that extends the BA format (which he calls *varint-30*) and reported significant performance gains obtained at Google by adopting it. A similar but more general format was described by Westmann et al. [19] in the database context. Schlegel et al. [17] applied SIMD-based decompression algorithms to a specialized version of Westmann's format that coincides exactly with group varint, but under the name *k-wise null suppression.* While they do not precisely describe the decoding and table generation, we believe that their algorithms are special cases of the generalized algorithms we describe in Section 6. All of these encodings fall into the taxonomy defined in this paper.

Büttcher et al. [4] have compared performance of several encoding techniques using posting lists from the GOV2 corpus on a particular query sample. They originally reported vbyte being the fastest at decoding, with Simple-9 being second. They recently updated their analysis [5] to include the group varint format, reporting that it outperforms both vbyte and Simple-9. Our experiments (presented in Section 7) show that the SIMD techniques described in this paper significantly outperform all of these in decoding speed.

3. BYTE-ORIENTED ENCODINGS

Encoding formats are generally distinguished by the granularity of their representation. We focus on encodings satisfying the following definition.

Definition 1. *We call an encoding* byte-oriented *if it satisfies the following conditions:*

1. *All significant bits of the natural binary representation are preserved.*
2. *Each byte contains bits from only one integer.*
3. *Data bits within a single byte of the encoding preserve the ordering they had in the original integer.*
4. *All bits from a single integer precede all bits from the next integer.*

A byte-oriented encoding is fixed-length *if every integer is encoded using the same number of bytes. Otherwise it is* variable-length.

Since variable-length byte-oriented formats must encode the length of the encoded data, they vary along the following three dimensions:

- The length can be expressed in binary or unary.
- The bits representing the length can be stored adjacent to the data bits of the corresponding integer so that some data bytes contain both data and length information; alternatively, lengths of several integers can be grouped together into one or more bytes distinct from the bytes containing data bits.
- If the length is represented in unary, the bits of the unary representation may be packed contiguously, or split across the bytes of the encoded integer.

It is evident that for byte-oriented formats, the natural unit of length is a byte. We call the set of bits used to represent the length the *descriptor*, since it describes how the data bits are organized.

Table 1: Nomenclature of Byte-Oriented Integer Encoding Formats

Descriptor Arrangement	Descriptor Length Encoding	Abbreviation	Names in the Literature
split	unary	varint-SU	v-byte [6], vbyte [4], varint [8], VInt [3], VB [13] (earlier references [14], [7] do not name the format)
packed	unary	varint-PU	none (format introduced in this paper)
group	unary	varint-GU	none (format introduced in this paper)
split	binary	varint-SB	none
packed	binary	varint-PB	BA [10], varint30 [8]
group	binary	varint-GB	group varint [8], k-wise (k=4) null suppression [17]

We assume that each encoded integer requires at least one byte, so both binary and unary descriptors can represent the length n by recording the value $n - 1$. This reduces the number of bits required to represent a given length.[1]

The dimensions listed above provide the basis of a taxonomy of byte-oriented encoding formats for integers that can be encoded in four bytes or less. Selecting one of the possible options for each dimension determines a position in the taxonomy. This taxonomy, shown in Table 1, provides a unifying nomenclature for the encoding formats, several of which have been described previously under various names.

For example, Grossman's BA format becomes *varint-PB* in our taxonomy, since it is a *variable*-length encoding of *integers* with descriptor bits *packed* together and representing the length in *binary.*

For the unary formats, we follow the natural convention where the quantity is represented by the number of consecutive 1 bits, followed by a terminating 0. We start from the least significant bit. Thus 0111 represents the number 3.

Accordingly, the vbyte encoding may be viewed as representing the length − 1 of the encoded representation in the sequence of continuation bits. For example, a three-byte integer encoding would look like this:

```
1xxxxxxx
1xxxxxxx
0xxxxxxx
```

Notice that the leading bits form the unary number 2, representing the length 3. Thus we call this representation *varint-SU*, since it is a *variable*-length representation of *integers* with length information *split* across several bytes and represented in *unary.* While unary length representation has been widely used in bit-oriented encodings, for byte-oriented encodings the concept of continuation bits obscured their interpretation as unary lengths.

If binary length descriptors are used, the descriptor length must be fixed in advance, or additional metadata would be required to store the length of the descriptor itself. For this reason, all binary formats in the taxonomy use fixed-length descriptors of 2 bits per integer. Furthermore, since splitting a fixed-length k-bit binary descriptor (one bit per byte) results in a byte-oriented integer encoding that requires at least k bytes, the split binary encoding format does not offer a competitive compression rate and we do not consider it further.

There are also additional variations of some of these formats. Bytes of the encoded data may be stored in little-endian or big-endian order; descriptor bits may be stored in the least significant or most significant bits. While these choices are sometimes described as arbitrary conventions, in practice there are efficiency considerations that make certain variants attractive for certain machine architectures. For example, in varint-SU, representing termination as 0 in the most significant bit allows the common case of a one-byte integer to be decoded without any shifts or masks. While traditional decoding algorithms run more efficiently when the representation preserves native byte ordering, the performance of the SIMD algorithms presented in Section 6 does not depend on the ordering. Without loss of generality, for the remainder of the paper we restrict our attention to little-endian encodings.

The byte-oriented encoding taxonomy suggests two encodings, *varint-PU* and *varint-GU*, that, to our knowledge, have not been previously described.

Varint-PU is similar to varint-SU, but with the descriptor bits packed together in the low-order bits of the first byte rather than being split across all bytes. (The choice of low-order bits to hold the descriptor is appropriate for little-endian encodings on little-endian architectures so that all data bits for one integer are contiguous. For the same reason, on big-endian architectures placing the descriptor in the high-order bits and using big-endian encoding is more efficient to decode.) The compression rate of varint-PU is the same as that of varint-SU, since the bits in each encoded integer are identical but rearranged. The decoding performance of varint-PU using unaligned reads, masks, and shifts in a table-driven algorithm similar to that for varint-PB is faster than traditional varint-SU, but significantly slower than the group formats such as varint-GU, described in detail in the next section.

4. THE GROUP ENCODING FORMATS

Within our taxonomy, encoding formats that group several integers together provide opportunities for exploiting SIMD parallelism. These encodings satisfy the following important property.

Definition 2. *We call a byte-oriented encoding* **byte-preserving** *if each byte containing significant bits in the original (unencoded) integer appears without modification in the encoded form.*

Neither split nor packed formats satisfy this property, since the descriptor bits are intermingled with data bits in some bytes. The separation of descriptor bytes from data

[1] Storing length as n would allow the length zero to represent an arbitrary constant with zero data bytes. Such an encoding, however, does not in general satisfy the first property of Definition 1.

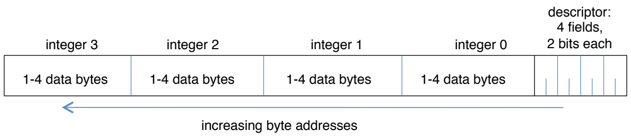

Figure 3: The varint-GB format.

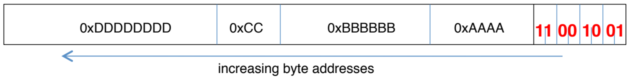

Figure 4: Storing the four integers `0xAAAA`, `0xBBBBBB`, `0xCC`, `0xDDDDDDDD` in varint-GB format. The value of each pair of bits in the descriptor is one less than the length of the corresponding integer. Byte addresses increase from right to left, matching the order of increasing bit significance. The order of pairs of bits in the descriptor matches the order of the integers.

bytes in group formats allows for more efficient decoding. It facilitates the use of tables to simplify the decoding process and avoids bitwise manipulations that are required to eliminate interspersed descriptor bits. In particular, we shall see in Section 6 that byte-preserving encodings are especially amenable to decoding with the SIMD techniques described in this paper.[2]

There are two classes of group formats, group binary (varint-GB) and group unary (varint-GU).

In the *varint-GB* format (called group varint in [8] and k-wise null supression (with $k = 4$) in [17]) a group of four integers is preceded by a descriptor byte containing four 2-bit binary numbers representing the lengths of the corresponding integers. Figure 3 illustrates this format for one such group. The actual number of bytes in a group may vary from 4 to 16. Figure 4 shows how the four hexadecimal numbers `0xAAAA`, `0xBBBBBB`, `0xCC`, `0xDDDDDDDD` would be represented. The four integers require, correspondingly, 2 bytes, 3 bytes, 1 byte, and 4 bytes. For each integer, its length n is represented in the descriptor by the 2-bit binary value $n - 1$. Therefore, the descriptor byte contains the values 01, 10, 00, and 11 respectively. To maintain a consistent order between descriptor bits and data bytes, we store the first binary length in the least significant bits, and so on. Thus the descriptor byte for these four integers is 11001001.

Varint-GB operates on a fixed number of integers occupying a variable number of bytes, storing their lengths in binary. In contrast, the varint-GU format operates on a fixed number of bytes encoding a variable number of integers, storing their lengths in unary.

Varint-GU groups 8 data bytes together along with one descriptor byte containing the unary representations of the lengths of each encoded integer. The 8 data bytes may encode as few as 2 and as many as 8 integers, depending on their size. The number of zeros in the descriptor indicates the number of integers encoded. This format is shown in Figure 5. The block size of 8 is the minimal size that can use every bit of the descriptor byte; larger multiples of 8 are possible, but did not improve performance in our experiments.

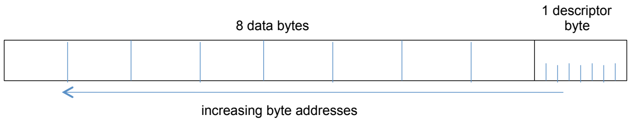

Figure 5: The varint-GU format.

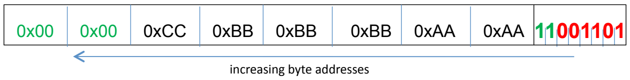

Figure 6: Storing the three integers (`0xAAAA`, `0xBBBBBB`, `0xCC`) in varint-G8IU format. The descriptor bits express the unary lengths of the integers. Since the next integer `0xDDDDDDDD` in our example does not fit in the data block, the block is left incomplete and padded with 0s, while the descriptor is padded with 1s.

Since not every group of encoded integers fits evenly into an 8-byte block, we have two variations of the encoding: incomplete and complete.

In the *incomplete* block variation, which we call *varint-G8IU*, we store only as many integers as fit in 8 bytes, leaving the data block incomplete if necessary.[3] The remaining space is padded with zeros, but is ignored on decoding.[4] When there is no additional integer to decode, the final (most significant) bits of the descriptor will be an unterminated sequence of 1 bits.

An example is shown in Figure 6. We use the same four integers `0xAAAA`, `0xBBBBBB`, `0xCC` and `0xDDDDDDDD` to illustrate. Encoding these values requires 10 bytes, but we have only 8 bytes in the block. The first three integers fit into the block using 6 bytes, leaving 2 bytes of padding. The final integer `0xDDDDDDDD` is left for the next block (not shown). The descriptor contains the three unary values 01, 011, and 0, and two padding bits 11. These are arranged in the same order as the integers, giving the descriptor a binary value of 11001101.

In the *complete* block variation, which we call *varint-G8CU*, we always fill all eight bytes in a data block.[5] As before, the number of zero bits in the descriptor indicates the number of complete integers encoded. In situations where an integer exceeds the remaining space in the current block, as much of that integer as fits is placed in the current block. The remaining bytes of that integer are carried over into the

[2] Note that what Anh [1] calls *word-aligned* is neither byte-oriented nor byte-preserving as defined in this paper.

[3] In our notations for the encoding, the number 8 represents the size of the data block.

[4] There is also a variation of this encoding format that uses variable size data blocks and avoids padding. Its performance characteristics are between those of varint-G8IU and varint-G8CU described later.

[5] As before, the number 8 represents the size of the data block.

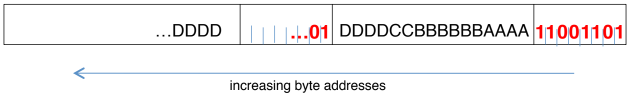

Figure 7: Storing four integers (`0xAAAA`, `0xBBBBBB`, `0xCC` and `0xDDDDDDDD`) in varint-G8CU format. The last two bytes of the fourth integer carry over to the subsequent data block, and its descriptor bits carry over to the subsequent descriptor byte.

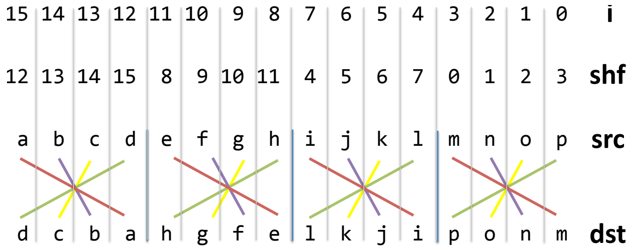
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 i
12 13 14 15 8 9 10 11 4 5 6 7 0 1 2 3 shf
a b c d e f g h i j k l m n o p src
d c b a h g f e l k j i p o n m dst

Figure 8: Using the `PSHUFB` instruction to reverse the byte order of four integers in parallel. The `shf` vector determines the shuffle sequence used to transform `src` to `dst`.

next data block. Similarly, the corresponding descriptor bits are carried over to the next block's descriptor byte.

An example is shown in Figure 7. Again we use the same four integers `0xAAAA`, `0xBBBBBB`, `0xCC` and `0xDDDDDDDD`. The first three integers and the corresponding descriptor bits are stored exactly as in varint-G8IU. However, varint-G8CU handles the fourth integer differently. Its first two bytes are placed in the first data block, filling it entirely, and the remaining two bytes go into the following block. The two descriptor bits corresponding to these last two bytes go into the next block's descriptor byte. Although spread across two descriptor bytes, the unary value of the descriptor bits for this fourth integer still equals $n-1$, where n is the length of the encoded integer.

5. SIMD ON INTEL/AMD

Facilities for fine-grained parallelism in the SIMD (Single Instruction, Multiple Data) paradigm are widely available on modern processors. They were originally introduced into general-purpose processors to provide vector processing capability for multimedia and graphics applications. Although SIMD instructions are available on multiple platforms, we restricted our focus to Intel-compatible architectures [12] implemented in current Intel processors in extensive use in many data centers, as well as forthcoming AMD processors.

In these architectures, a series of SIMD enhancements have been added over time. Among the current SIMD capabilities are 16-byte vector registers and parallel instructions for operating on them.

The `PSHUFB` instruction, introduced with SSSE3 in 2006, is particularly useful.[6] It performs a permutation ("shuffle") of bytes in a vector register, allowing the insertion of zeros in specified positions. `PSHUFB` has two operands, a location containing the data and a register containing a shuffle sequence. If we preserve the original value of the data operand, we can view `PSHUFB` as transforming a source sequence of bytes *src* to a destination sequence *dst* according to the shuffle sequence *shf*, implementing the following algorithm:

```
for $0 \leq i < 16$ parallel do
    if $shf[i] \geq 0$ then
        $dst[i] \leftarrow src[shf[i] \bmod 16]$
    else
        $dst[i] \leftarrow 0$
end
```

In other words, the ith value in the shuffle sequence indicates which source byte to place in the ith destination byte. If the ith value in the shuffle sequence is negative, a zero is placed in the corresponding destination byte.

The example illustrated in Figure 8 shows how `PSHUFB` can be used to reverse the byte order of four 32-bit integers at once.

6. SIMD DECODING

Since byte-preserving formats remove leading zero bytes while retaining the significant bytes intact, the essence of decoding is reinserting the zero bytes in the right places. Our discovery is that we can make `PSHUFB` do virtually all of the work for many formats. We construct a shuffle sequence by inserting -1s in a sequence $\{0, 1, 2, 3, ...\}$. With this sequence, the `PSHUFB` instruction will copy the significant data bytes while inserting the missing zeros.

An example of using `PSHUFB` to decode varint-G8IU is shown in Figure 9. This is the same data represented in Figure 6.

For a given format, we can precompute what the correct shuffle sequence is for a particular data block and its corresponding descriptor byte. For all possible values of the descriptor (and sometimes additional state) we build a table of any shuffle sequence that might be needed at decode time.

The table entries also contain a precomputed offset. For the varint-GB format, the offset indicates how many bytes were consumed to decode 4 integers; it always outputs 16 bytes. For the varint-GU formats, the offset indicates how many integers were decoded; it always consumes 8 bytes.

Table construction occurs only once, while table lookup occurs every time a group is decoded.

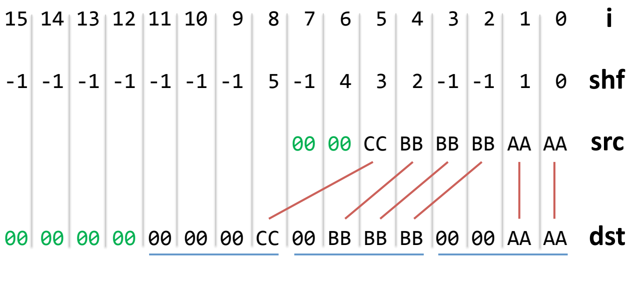
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 i
-1 -1 -1 -1 -1 -1 -1 5 -1 4 3 2 -1 -1 1 0 shf
00 00 CC BB BB BB AA AA src
00 00 00 00 00 00 00 CC 00 BB BB BB 00 00 AA AA dst

Figure 9: Using the `PSHUFB` instruction to decode the varint-G8IU format.

[6] A similar instruction, `vperm`, is part of the AltiVec/VMX instruction set for the PowerPC processor family.

Given the availability of these tables, the general strategy of all the decodings is:

1. read a chunk of data and its corresponding descriptor;
2. look up the appropriate shuffle sequence and offset from the table;
3. perform the shuffle;
4. write the result;
5. advance the input and output pointers.

This approach allows us to decode several integers simultaneously with very few instructions. It requires no conditionals, and thus avoids performance penalties due to branch misprediction. Two techniques make this possible. First, the logical complexity has been shifted from the code to the table. Second, the algorithm always reads and writes a fixed amount and then relies on the table to determine how much input data or output data it has actually processed.[7]

Data blocks are not aligned on any fixed boundary. We depend on the ability of the CPU to perform unaligned reads and writes efficiently, and we have observed this to be true on modern Intel processors.

6.1 Details of the Decoding Algorithm

All of the group formats can be decoded using the generalized algorithm shown in Algorithm 1.

Algorithm 1: decodeBlock

```
Decodes a block of data using SIMD shuffle.
input  : $src, dst, state$
output: $src, dst, state$
begin
    $data \leftarrow \text{read}(src + 1, 16)$
    $entry \leftarrow table_{format}[desc, state]$
    $shf \leftarrow \text{shuffleSequence}(entry)$
    $\text{shuffleAndWrite}(data, shf, dst)$
    $src \leftarrow src + \text{inputOffset}(entry)$
    $dst \leftarrow dst + \text{outputOffset}(entry)$
    return $src, dst, \text{nextState}(entry)$
end
```

Because this algorithm constitutes the inner loop of the decoding process, it is essential to inline the implementation to avoid function call overhead. The algorithm takes three inputs:

- *src* – a pointer to the input byte stream of encoded values
- *dst* – a pointer to the output stream of integers in case of varint-GB, and varint-G8IU; in the case of varint-G8CU, *dst* is a pointer to an output stream of bytes, since decoding a block of varint-G8CU may result in writing a portion of a decoded integer.
- *state* – auxiliary state. This is required only for varint-G8CU, where it is an integer i, with $0 \leq i < 4$ indicating the number of bytes modulo 4 of the last integer written.

The algorithm reads encoded values from the input stream, outputs decoded integers to the output stream and returns as its result the new positions of *src*, *dst*, and the updated *state*.

We always read 16 bytes, the size of the vector register used by the PSHUFB operation. The number of bytes corresponding to a single byte descriptor is 8 for the unary formats and at most 16 for the binary format. While it is possible to read only 8 bytes for the unary formats, in our evaluation we found that doing so did not improve performance, and in fact made the implementation slightly slower.

A different table is used for each format. There is a table entry corresponding to each possible descriptor value and state value. The table has 256 entries for the varint-GB and varint-G8IU formats. For varint-G8CU format, the table has $4 \times 256 = 1024$ entries, because we have an entry for each descriptor and state pair, and the state is an integer i, with $0 \leq i < 4$.

Each table entry logically contains four things:

- a shuffle sequence
- an input offset
- an output offset
- the state value to use for the subsequent block

The shuffleSequence, inputOffset, outputOffset, and nextState functions are accessors for these fields. For some of the formats, some of these values are constant over all entries in the table, and are not stored explicitly; the accessors simply return constant values.

The shuffleAndWrite operation uses the PSHUFB operation with the provided shuffle sequence to expand the 16 bytes of data, inserting zeros into the correct positions. It then writes its result to the destination.

In the varint-GB case, the shuffle sequence is a 16-byte sequence describing a single PSHUFB operation. A single PSHUFB is sufficient because the group always contains four encoded integers, and thus the output never exceeds 16 bytes.

For decoding the varint-GU formats, the shuffle sequence is a 32-byte sequence specifying two PSHUFB operations. The second PSHUFB is required for the unary formats because an 8-byte data block may encode up to 8 integers, which can expand to 32 bytes. The output of the first PSHUFB is written to locations beginning at *dst*, and the output of the second PSHUFB to locations beginning at $dst + 16$. To avoid conditionals, the second shuffle is always performed, even when the output does not exceed 16 bytes. Since PSHUFB rearranges the register in place, the corresponding register needs to be reloaded with the original data before the second PSHUFB.

For unary formats, the input offset, by which we increment the *src*, is always 8 bytes. For varint-G8IU, the output offset measured in units of decoded integers varies between 2 and 8, except for the last block of a sequence, which may contain only 1 integer. For varint-G8CU, decoding one block may result in writing a portion of a decoded integer, so the output is a byte stream and the offset is measured in byte units. It varies between 8 and 32 bytes, except for the last block of the sequence which may output only 1 byte.

In the case of varint-GB, the output offset is always a

[7] This requires that the input and output buffers always have at least this amount available to read or write.

constant 4 integers.[8] The input offset varies between 4 and 16 bytes.

For all of the encodings, the input offset needs to account for the additional one byte of the descriptor as well. All variable offsets are precomputed and stored in the format table.

For the varint-G8CU format, the table also contains the new state information indicating the number of bytes in the last integer to be used to decode the subsequent block.

6.2 Building the Tables

For each of the group formats, the decoding table used by Algorithm 1 is constructed in advance. The construction process takes as input a descriptor byte value and a state value. It builds the shuffle sequence for the entry and computes the input offset, output offset, and next state (unless they are constant for the format).

We assume we deal only with valid descriptor values, those which could actually arise from encoding. For varint-GB, all possible byte values are valid. For the group unary formats, a descriptor is valid if and only if the distance between consecutive zero bits does not exceed 4.

The algorithms for constructing shuffle sequences, offset values, and the next state value depend on the following abstract functions:

- $\mathrm{num}(desc)$ gives the number of integers whose encoding is completed in the group described by the descriptor value *desc*. For varint-GB this is always 4. For the group unary formats, this value is the number of 0 (termination) bits in *desc*.

- $\mathrm{len}(desc, i)$ gives the length of the ith integer in the group, for each i, $0 \leq i < \mathrm{num}(desc)$. This is the length determined by the ith individual bit pair in *desc* for varint-GB, or the ith unary value in *desc* for the unary formats.

- $\mathrm{rem}(desc)$ gives the number of bytes modulo 4 in the last encoded integer in the group. This is needed only for varint-G8CU, where it is equal to the number of leading 1s in the descriptor *desc*. For the other formats it is always zero.

Again, the basic idea in constructing a shuffle sequence is to insert -1s in a sequence $\{0, 1, 2, 3, ...\}$ representing the byte positions in one block of the source data being decoded. The resulting shuffle sequence is used by the `PSHUFB` instruction to copy the significant data bytes while inserting the missing leading zeros. The details of the construction are shown in Algorithm 2. The algorithm takes two inputs:

- *desc* the descriptor value

- *state* the number of bytes modulo 4 written from the last integer in the prior group.[9]

The algorithm produces one output, *shf*, the shuffle sequence to be used for the given descriptor and state. The first loop iterates over every completed integer in the group corresponding to the given descriptor. For each completed integer in the group, the inner loop sets the shuffle sequence to move the encoded bytes from the source of the shuffle operation, inserting -1s to produce the leading zeros necessary to complete the decoded integer. Here the variable j advances over the source data positions in the data block, while the variable k advances over the positions in the shuffle sequence, which correspond to destination positions of the shuffle operation.

The concluding loop only executes for varint-G8CU. It sets the remainder of the shuffle sequence to transfer encoded bytes from the source for the last incomplete integer in the group.

Algorithm 2: constructShuffleSequence

```
input  : $desc, state$
output: $shf$
begin
    $j, k \leftarrow 0$
    $s \leftarrow 4 - state$
    for $0 \leq i < \mathrm{num}(desc)$ do
        for $0 \leq n < s$ do
            if n $< \mathrm{len}(desc, i)$ then
                $shf[k] \leftarrow j$
                $j \leftarrow j + 1$
            else
                $shf[k] \leftarrow -1$
            end
            $k \leftarrow k + 1$
        end
        $s \leftarrow 4$
    end
    for $0 \leq n < \mathrm{rem}(desc)$ do
        $shf[k] \leftarrow j$
        $j \leftarrow j + 1$
        $k \leftarrow k + 1$
    end
    return $shf$
end
```

Computing input offsets is easy. For the unary formats the input offset is always 9; we always consume a block of 8 bytes of data and 1 descriptor byte. For the group binary format varint-GB, the input offset for a given descriptor *desc* is

$$1 + \sum_{i=0}^{3} \mathrm{len}(desc, i)$$

which is the sum of the lengths of the integers in the group plus 1 for the descriptor byte.

The output offset for varint-GB and varint-G8IU is equal to $\mathrm{num}(desc)$ integers (which is always 4 for varint-GB). The output offset is

$$4 \cdot \mathrm{num}(desc) - state + \mathrm{rem}(desc)$$

for the varint-G8CU format.

The state value for the subsequent block is always 0 for varint-GB and varint-G8IU (and the state can be ignored for these formats). For varint-G8CU it is $\mathrm{rem}(desc)$.

[8] The varint-GB format requires auxiliary information to deal with sequences of length not divisible by 4. This may be done using length information stored separately or the convention that zero values do not appear in the sequence, so terminal zeros can be ignored.

[9] For varint-GB and varint-G8IU, the value of state is always zero, since only complete integers are written in a given data block in these formats.

Table 2: Decoding rates in millions of integers per second; larger is better

encoding	algorithm	Wikipedia	Reuters	GOV2
varint-SU	traditional	424	516	491
varint-SU	SIMD	547	640	477
varint-GB	mask table	766	831	763
varint-GB	SIMD	1159	1276	1024
varint-G8IU	SIMD	**1321**	**1518**	**1059**
varint-G8CU	SIMD	1231	1398	1033

Table 3: Compression ratios; smaller is better

encoding	Wikipedia	Reuters	GOV2
varint-SU	0.34	0.30	0.32
varint-GB	0.38	0.35	0.37
varint-G8IU	0.37	0.33	0.35
varint-G8CU	0.36	0.33	0.34

6.3 Applying SIMD to varint-SU

Although varint-SU is not a byte-preserving encoding, Algorithm 1 can be applied as a component for decoding it. By efficiently gathering the descriptor bits from each byte into a single value for a table lookup, we can treat a sequence of 8 consecutive varint-SU-encoded bytes almost as if they were a varint-GU-encoded block. The Intel instruction `PMOVMSKB`, which gathers the most-significant bits from 8 bytes into a single byte, provides the needed functionality. After applying a shuffle as in Algorithm 1, the descriptor bits must be "squeezed out" to finish the decoding; this removal of the interspersed descriptor bits requires masks and shifts. Despite this additional step, the SIMD implementation still outperforms the traditional method in most cases as shown in Section 7.

7. EVALUATION

We used three corpora in our evaluation: Wikipedia [20] (6M documents[10], 27 GB), Reuters RCV1 [16] (0.8M documents, 2.5 GB), and GOV2 [15] (25M documents, 426 GB). To satisfy the resource constraints of our test environment, we randomly sampled 50% of the Wikipedia documents and 15% of GOV2. We removed all XML/HTML markup and five common stopwords (`a, an, and, of, the`) and applied stemming to conflate singular and plural nouns.

The C++ implementation was compiled using gcc 4.5.1.[11] Measurements were done using a single-threaded process on an Intel Xeon X5680 processor (3.3GHz, 6 cores, 12 MB Intel Smart Cache shared across all cores), with 24 GB DDR3 1333 MHz RAM. Measurements are done with all of the input and output data in main memory.

Decoding speed results are shown in Table 2. For each corpus, this table shows the decoding speed measured in millions of integers per second; the fastest result is shown in boldface. In every case, a SIMD algorithm strongly outperformed a conventional algorithm; in all cases, the implementation of our varint-G8IU format was fastest. The "traditional" implementation of varint-SU shown in the table is our best implementation for this encoding using traditional techniques. The "mask table" implementation of varint-GB is our implementation of the technique described in Dean [8].

Compression ratios are shown in Table 3. Here the compression ratio indicates the ratio between the bytes required for the integers encoded in the format and their original size of 4 bytes each. Compression ratios depend only on the encoding and not on the implementation.

7.1 Comparison with Other Evaluations

There seems to be no standard benchmark for measuring integer decoding implementations in the information retrieval field. Even with conventional test corpora, there are many variations possible in producing the posting lists. Evaluation methods are also not standardized. Some research on compression only reports compression rate but not speed. Speed is reported using different metrics and data.

Büttcher et al. [4] start with the GOV2 corpus and the 10000 queries from the efficiency task of the TREC Terabyte Track 2006. They indicate they used components of the Wumpus search engine to index the GOV2 corpus, and then decoded the posting lists for non-stopword terms contained in the query set. They do in-memory measurements, but also compute a "cumulative overhead" for decoding and disk I/O by estimating the I/O time based on compression rate.

Schlegel et al. [17] use 32MB synthetic data sets constructed by sampling positive 32-bit integers from the Zipf distribution with different parameter values. They measure the amount of uncompressed data that can be processed per second.

Dean [8] uses millions of integers decoded per second as a performance metric (as we do), but he does not provide details on the data or evaluation method used in his measurements.

We measured performance on standard corpora decoding every posting list once, repeating the test to achieve stable timing. While it is easy to reproduce, this method does not account for different term distributions in queries. To account for different frequencies of terms in real queries, one needs a representative query mix. Since GOV2 and Reuters

[10] This counts all documents from the English Wikipedia dump except empty documents and redirect pages.

[11] It uses the GCC intrinsics `__builtin_ia32_pshufb128`, `__builtin_ia32_loaddqu`, and `__builtin_ia32_storedqu` to invoke the `PSHUFB` and unaligned load and store instructions.

RCV1 are not used in real world search tasks, such a mix is not available.

Creation of a standard benchmark containing several sequences of integers with distinct but representative statistical characteristics would allow meaningful comparisons of different implementations.

8. CONCLUSIONS AND FUTURE WORK

We discovered a taxonomy for variable-length integer encoding formats. This led us to identify some new encodings that offer advantages over previously known formats. We identified the *byte-preserving* property of encoding formats which makes them particularly amenable to parallel decoding with SIMD instructions. The SIMD-based algorithms that we developed outperformed the traditional methods by 300%, and the best previously published methods by over 50%. Furthermore the new group unary formats offer better compression than the group binary format on all of the corpora tested.

Schlegel et al. [17] also reported success applying SIMD to Elias γ [9], which is not aligned on byte boundaries at all. Further investigations are needed to see whether SIMD techniques can be similarly applied to other encodings which are not aligned on byte boundaries, such as Simple-9 [2], PForDelta [21] and VSEncoding [18].

We restricted our investigation to integers that can be encoded in four bytes or less. We believe, however, that some of the encodings that we introduced could be easily extended to larger values.

SIMD instructions, a powerful but under-utilized resource, offer the opportunity for significant performance improvements in modern search engines.

Acknowledgments

We thank Thomas London for suggestions on a draft of this paper and Bill Stasior for supporting this research.

References

[1] V. N. Anh. *Impact-Based Document Retrieval.* PhD thesis, University of Melbourne, April 2004.

[2] V. N. Anh and A. Moffat. Inverted index compression using word-aligned binary codes. *Information Retrieval*, 8(1):151–166, 2005.

[3] Apache Software Foundation. Lucene 1.4.3 documentation. `http://lucene.apache.org/java/1_4_3/fileformats.html`, 2004.

[4] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines.* MIT Press, Cambridge, MA, 2010.

[5] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. Information retrieval: Implementing and evaluating search engines, addenda for chapter 6: Index compression. `http://www.ir.uwaterloo.ca/book/addenda-06-index-compression.html`, 2010.

[6] W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice.* Pearson Education, Boston, 2010.

[7] D. Cutting and J. Pedersen. Optimizations for dynamic inverted index maintenance. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, pages 405–411, New York, NY, USA, 1990. ACM.

[8] J. Dean. Challenges in building large-scale information retrieval systems. Keynote, WSDM 2009, `http://research.google.com/people/jeff/WSDM09-keynote.pdf`, February 2009.

[9] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21(2):194–203, 1975.

[10] D. A. Grossman. *Integrating Structured Data and Text: A Relational Approach.* PhD thesis, George Mason University, 1995.

[11] H. S. Heaps. Storage analysis of a compression coding for a document database. *INFOR*, 10(1):47–61, February 1972.

[12] Intel Corporation. *Intel 64 and IA-32 Architectures Software Developers Manual.* Intel Corporation, Santa Clara, California, USA, September 2010. Version 37.

[13] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[14] MIDI Manufacturers Association. *MIDI 1.0 Specification*, 1982-2001.

[15] NIST. GOV2 collection. `http://ir.dcs.gla.ac.uk/test_collections/`, 2010.

[16] Reuters. Reuters RCV1 Corpus. `http://trec.nist.gov/data/reuters/reuters.html`, 2010.

[17] B. Schlegel, R. Gemulla, and W. Lehner. Fast integer compression using SIMD instructions. In *Proceedings of the Sixth International Workshop on Data Management on New Hardware (DaMoN 2010)*, Indianapolis, Indiana, June 7 2010.

[18] F. Silvestri and R. Venturini. VSEncoding: efficient coding and fast decoding of integer lists via dynamic programming. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1219–1228, New York, NY, USA, 2010. ACM.

[19] T. Westmann, D. Kossmann, S. Helmer, and G. Moerkotte. The implementation and performance of compressed databases. *SIGMOD Rec.*, 29:55–67, September 2000.

[20] Wikimedia Foundation. Wikipedia database download (english). `http://en.wikipedia.org/wiki/Wikipedia:Database_download`, September 2010.

[21] M. Zukowski, S. Héman, N. Nes, and P. A. Boncz. Super-scalar RAM-CPU cache compression. In *Proceedings of the 22nd International Conference on Data Engineering*, ICDE '06, pages 59–, Washington, DC, USA, 2006. IEEE Computer Society.