

SHORT COMMUNICATION

Base64 encoding and decoding at almost the speed of a memory copy

Wojciech Muła | Daniel Lemire*

¹TELUQ, Université du Québec, Quebec, Canada

Correspondence

*D. Lemire, Université du Québec (TELUQ), 5800, Saint-Denis street, Montreal (Quebec) H2S 3L5, Canada.
Email: daniel.lemire@teluq.ca

Summary

Many common document formats on the Internet are text-only such as email (MIME) and the Web (HTML, JavaScript, JSON and XML). To include images or executable code in these documents, we first encode them as text using base64. Standard base64 encoding uses 64 ASCII characters: both lower and upper case Latin letters, digits and two other symbols. We show how we can encode and decode base64 data at nearly the speed of a memory copy (`memcpy`) on recent Intel processors, as long as the data does not fit in the first-level (L1) cache. We use the SIMD (Single Instruction Multiple Data) instruction set AVX-512 available on commodity processors. Our implementation generates several times fewer instructions than previous SIMD-accelerated base64 codecs. It is also more versatile, as it can be adapted—even at runtime—to any base64 variant by only changing constants.

KEYWORDS:

Binary-to-text encoding, Vectorization, Data URI, Web Performance

1 | INTRODUCTION

Base64 formats represent arbitrary binary data as ASCII text: e.g., the ubiquitous MIME email protocol¹ require that arbitrary binary attachments be encoded as ASCII, typically using base64. In a similar manner, binary data (e.g., images, WebAssembly code) can also be included inside text-only resources like XML, JavaScript or HTML documents. Base64 is part of the standard library of many popular programming languages (Java, Go, JavaScript, Swift), it is used in many important database systems. Some database systems even store binary data as base64 (MongoDB, Elasticsearch, Amazon SimpleDB and DynamoDB). Crane and Line² observe that decoding all this base64 data can be a performance concern.

Commodity processors support single-instruction-multiple-data (SIMD) instructions. The SIMD extensions AVX-512 available on recent Intel processor operate on 512-bit registers. Hence, we can compare two strings of 64 characters using a single instruction. Software that uses fewer instructions³ tend to be faster and use less energy.

We refer to algorithms designed to benefit from SIMD instructions as being vectorized. We can reasonably expect that as new processor architectures include wider SIMD registers, corresponding vectorized algorithms can run faster. A doubling of the register width might almost double the performance—everything else being equal. In earlier⁴ work, we showed that we could greatly increase base64 encoding and decoding speeds by using 256-bit registers available in AVX and AVX2 instruction set extensions. Such a vectorized approach was later adopted by the PHP interpreter and other systems.

Maybe surprisingly, we find that we can more than double the speed by using the new 512-bit instructions: for example, we achieve a five-fold reduction in the number of instructions during decoding by moving from AVX2 to AVX-512. For inputs that

TABLE 1 Base64 mapping between 6-bit values and ASCII characters.

value	ASCII	char	value	ASCII	char	value	ASCII	char	value	ASCII	char
0	0x41	A	16	0x51	Q	32	0x67	g	48	0x77	w
1	0x42	B	17	0x52	R	33	0x68	h	49	0x78	x
2	0x43	C	18	0x53	S	34	0x69	i	50	0x79	y
3	0x44	D	19	0x54	T	35	0x6a	j	51	0x7a	z
4	0x45	E	20	0x55	U	36	0x6b	k	52	0x30	0
5	0x46	F	21	0x56	V	37	0x6c	l	53	0x31	1
6	0x47	G	22	0x57	W	38	0x6d	m	54	0x32	2
7	0x48	H	23	0x58	X	39	0x6e	n	55	0x33	3
8	0x49	I	24	0x59	Y	40	0x6f	o	56	0x34	4
9	0x4a	J	25	0x5a	Z	41	0x70	p	57	0x35	5
10	0x4b	K	26	0x61	a	42	0x71	q	58	0x36	6
11	0x4c	L	27	0x62	b	43	0x72	r	59	0x37	7
12	0x4d	M	28	0x63	c	44	0x73	s	60	0x38	8
13	0x4e	N	29	0x64	d	45	0x74	t	61	0x39	9
14	0x4f	O	30	0x65	e	46	0x75	u	62	0x2b	+
15	0x50	P	31	0x66	f	47	0x76	v	63	0x2f	/

do not fit in the fastest CPU cache (L1), it is nearly exactly as fast to copy base64 data than to encode or decode it. The key to this greater- than-expected performance is an algorithmic redesign to benefit from the new instructions introduced in AVX-512.

2 | BASE64

The ASCII character set is made of 128 code points with byte values in $[0, 128)$; we can recognize a stream of ASCII characters by the fact the most significant bit of each byte is zero. Base64 code is made of 64 ASCII characters: all 26 letters (upper case and lower case), all ten digits and two other characters ('+', '/'). Each of these 64 characters correspond to a 6-bit unsigned integer value in the interval $[0, 64)$. Table 1 provides a bidirectional mapping between the integer values and ASCII characters. This map is invertible.

Given three arbitrary byte values spanning $3 \times 8 = 24$ bits, we can always represent them using four ASCII characters from the base64 set since $4 \times 6 = 24$ bits. During encoding, if the input is not divisible by three bytes, then one or two special padding character ('=') may be appended so that the length of the base64 code is divisible into blocks of four ASCII characters. Given three byte values s_1, s_2, s_3 , the base64 standard maps them bijectively to the four 6-bit values $s_1 \div 4, (s_2 \div 16) + (s_1 \times 16) \bmod 64, (s_2 \times 4) \bmod 64 + (s_3 \div 64)$ and $s_3 \bmod 64$. We can also use other lists of ASCII characters as in the base64url standard⁵ where the characters '+' and '/' are replaced by '-' (minus) and '_' (underline).

We can quickly encode and decode base64 strings using look-up tables. It is the approach taken by the Google Chrome browser. For example, we can map ASCII characters to integers in $[0, 64)$ as in Table 1, with a special value to identify ASCII characters outside of the base64 domain (for error detection). Given four bytes mapped to integer values in $[0, 64)$, a, b, c, d , we can retrieve that original three bytes as $(a \times 4) + (b \div 16), (b \times 16) \bmod 256 + (c \div 4), (c \times 64) \bmod 256 + d$.

3 | ALGORITHMIC DESIGN

We seek to encode to base64 and decode from base64 using as few instructions as possible. To this end, we load our data into 512-bit (64 bytes) registers and we use instructions that operate on full 512-bit registers. Because encoding *expands* the number of bytes, we need to consume 48 bytes if we want the encoder to produce 64 bytes with each iteration. Similarly, we want the decoder to consume 64 bytes and produce 48 bytes per iteration.

The AVX-512 instruction sets⁶ were first introduced by Intel in 2013. It consists of several distinct instructions sets, not all of which are included on all processors with AVX-512 support. Within AVX-512, we use instructions from the *Vector Byte Manipulation Instructions* (VBMI) set which we expect to be included in all new Intel microarchitectures (e.g., Cannon Lake, Cascade Lake, Cooper Lake, Ice Lake) for laptops and servers. AVX-512 instructions have been used to accelerate the implementation of algorithms in genomics⁷, machine learning⁸, databases^{9,10,11} as well as in high-performance computing.

3.1 | Encoding

If we omit load and store instructions, only two `vperm` and one `vpshl`, i.e. just three instructions, are necessary to encode 48 bytes into base64. In contrast, the previously best vectorized base64 encoder⁴ based on AVX2 and 256-bit registers uses 11 SIMD instructions to encode 24 bytes into base64. In other words, we achieve a seven-fold reduction in instruction count for a given number of bytes—it exceeds the reduction by a factor of two that we may expect from a doubling of the register size from 256 to 512 bits.

We use only two distinct AVX-512 instructions:

- The `vperm` instruction takes two 512-bit registers as inputs and treats them as arrays of 64 bytes. One register is an array of indexes (x), the other is a collection of values (a). The instruction computes the composition of the values and the indexes, ($a[x[0]], a[x[1]], \dots, a[x[63]]$), with the convention that only the six least significant digits of index values are used, the two most significant bits are silently ignored.

The characteristics of this instruction are particularly useful because base64 has exactly 64 ASCII output characters.

- The `vpshl` instructions takes two 512-bit registers as inputs. One register is made of eight 64-bit words, whereas the other register is an array of 64 bytes that are interpreted as shifts. The instruction executes a total of 64 shifts.

The instruction processes each 64-bit word, separately. Each such 64-bit word from the first register is paired with eight shift bytes from the second register.

Given a 64-bit word and a given shift byte, we take the least significant 6 bits of the shift byte (as a number between 0 and 64), rotate right the 64-bit word by this amount, and then output the least significant 8 bits.

We can separate the encoding task into two steps. Our goal is to map each sequence of three bytes s_1, s_2, s_3 to the four 6-bit values $s_1 \div 4, (s_2 \div 16) + (s_1 \times 16) \bmod 64, (s_2 \times 4) \bmod 64 + (s_3 \div 64)$ and $s_3 \bmod 64$.

1. We first want to map each block of three 8-bit values to a block of four 6-bit values. We plan to produce 64 ASCII characters so we consume only first 48 bytes of the register, ignoring the last 16 bytes of the 64-byte register.

We use the `vperm` instruction to map each sequence of three bytes (s_1, s_2, s_3) into a sequence of four bytes (s_2, s_1, s_3, s_2). Thus if we start with bytes values 0 to 63 (grouped in sets of three)

0	1	2	3	4	5
6	7	8	9	10	11
12	13	14	15	16	17
18	19	20	21	22	23
24	25	26	27	28	29
30	31	32	33	34	35
36	37	38	39	40	41
42	43	44	45	46	47
48	49	50	51	52	53
54	55	56	57	58	59
60	61	62	63		

we would finish, after the mapping, with byte values (grouped in sets of four)

1	0	2	1	4	3	5	4
7	6	8	7	10	9	11	10
13	12	14	13	16	15	17	16
19	18	20	19	22	21	23	22
25	24	26	25	28	27	29	28
31	30	32	31	34	33	35	34
37	36	38	37	40	39	41	40
43	42	44	43	46	45	47	46

As our example illustrates, the last sixteen values from the input are unused. Moreover, each block of three input bytes occupies 32 bits (4 bytes) in the shuffled register. To complete the encoding, we need to reorganize the bits in each 32-bit block to fit the base64 format.

To arrive at this desired bit layer, we apply the `vpmultishiftqb` instruction. Each 8-byte word is shifted eight times with the shift values

$$\begin{array}{cccc} 10, & 4, & 22, & 16, \\ 10 + 32, & 4 + 32, & 22 + 32, & 16 + 32. \end{array}$$

It is best to view our application of the `vpmultishiftqb` on an 4-byte basis even though it operates on 8-byte words. To see what these shifts do, let us consider any four packed bytes s_2, s_1, s_3, s_2 . We label the bit values from s_1 to a, s_2 to b, and s_3 to c. Similarly add another sequence of four bytes, this time with the labels d,e,f. We start with the following content

[bbbbbbbB|aaaaaaaA|cccccccC|bbbbbbbB|eeeeeeeE|dddddddD|fffffffF|eeeeeeeE].

We use the upper case letter to indicate the most significant bit of each byte value. We then get the following after applying the various right shifts (10, 4, ...) and selecting only the least significant eight bits:

10 : aaaaaAcc,
4 : bbbBaaaa,
22 : cCbbbbbb,
16 : cccccccC,
10+32: ddddDff,
4+32 : eeeEdddd,
22+32: fFeeeeee,
16+32: ffffffff.

We can then pack the result to determine the result of the `vpmultishiftqb` instruction:

[aaaaaAcc|bbbBaaaa|cCbbbbbb|cccccccC|ddddDff|eeeEddd|fFeeeeee|fffffffF].

Let us ignore the most significant two bits of each byte:

[aaaaaA--|bbbBaa--|cCbbbb--|cccccc--|ddddD--|eeeEdd--|fFeeee--|ffffff--].

If we ignore the two most significant bits of each byte, we see that starting from sequences of four byte values s_2, s_1, s_3, s_2 , we output the four byte values $s_1 \div 4, (s_2 \div 16) + (s_1 \times 16) \bmod 64, (s_2 \times 4) \bmod 64 + (s_3 \div 64)$ and $s_3 \bmod 64$. We do not have to explicitly reset these two bits because, in the upcoming second step, the `vpermB` instruction does it internally.

The `vpmultishiftqb` allows us to use fewer instructions. In our AVX2 algorithm⁴, this bit shifting requires as much as five instructions: two bitwise AND instructions, two variable shifts instructions and one bitwise OR instruction.

2. We then need to map the blocks of 6-bit values using the `vpermB` instruction. The 6-bit values are used as indexes while the values are the list of possible ASCII values as a 64-byte register:

ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789+/-.

We can merely replace a few characters in this register to support other base64 variants⁵ like `base64url`: in fact any 64-byte mapping is feasible, even if determined dynamically at runtime.

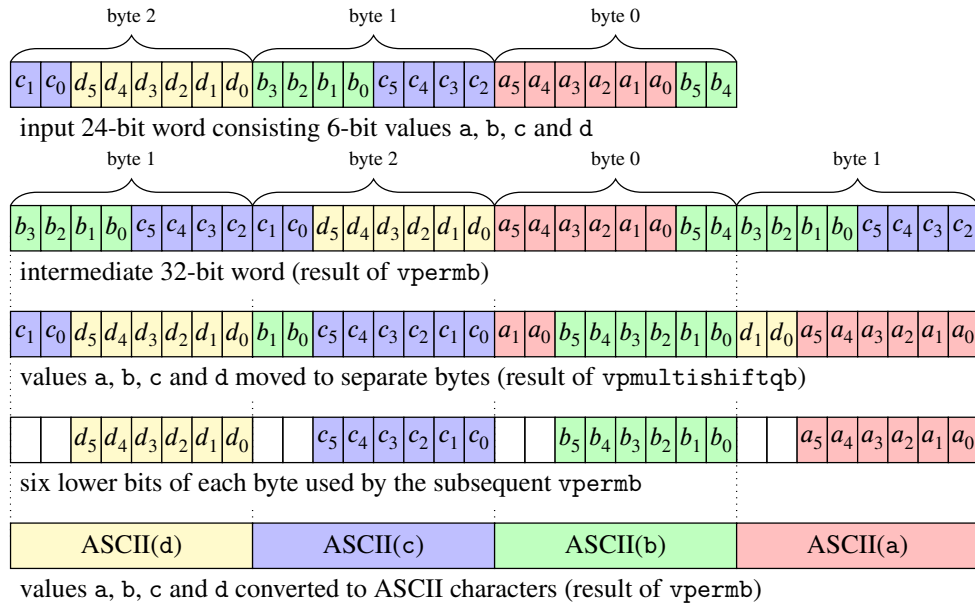


FIGURE 1 Vectorized encoding algorithm

We summarize the encoding process in Fig. 1. In practice, the input to be encoded may not be divisible by 48 bytes: we process any leftover bytes using a conventional code path.

3.2 | Decoding

One might expect decoding to be just like encoding but in reverse. However, we must validate that the input is proper base64 text. Though the encoding requires only three instructions per block of 64 bytes, we need five instructions to decode 64 bytes: `vpermi2b`, `vpternlogd`, `vpaddubsw`, `vpaddwd`, and `vperm`. A single call to an additional instruction (`vpmovb2m`) is needed, once per base64 stream. In contrast, the best 256-bit vectorized decoder⁴ uses 14 SIMD instructions to decode 32 input ASCII bytes into 24 output bytes. Thus we almost achieve a five-fold reduction in the instruction count for the same number of input bytes. Again, it is far better than the reduction by a factor of two that the wider registers would suggest.

In addition to the `vperm` instruction, already present in the encoder, we use the following five AVX-512 instructions:

- The `vpermi2b` instruction is a more powerful version of the `vperm` instruction. It takes three 512-bit registers. One of them (x) acts as an array of 64 byte-valued indexes, only the least significant 7 bits of each byte is considered, the most significant bit is ignored. Thus the register is treated as an array of integer indexes in the interval $[0, 128)$. The most significant bit in each index value of x is just ignored. The other two registers (a, b) form a 128-byte array used for lookup. When the index value is less than 64, one register is used, otherwise the other is used. Let `choose(i, x, a, b)` output $a[x[i]]$ when $x[i] < 64$ and $b[x[i] - 64]$, with the convention that $x[i]$ is a 6-bit value in $[0, 128)$. We can formally represent the result of the `vpermi2b` instruction as $(\text{choose}(0, x, a, b), \text{choose}(1, x, a, b), \dots, \text{choose}(64, x, a, b))$.
- The `vpmovb2m` instruction creates a 64-bit mask value made of all of the most significant bits of all 64 input bytes taken from a single 512-bit register. Thus to check that all byte values in a registers are 7-bit values—i.e. are smaller than 0x80—it suffices to check whether the result of the `vpmovb2m` instruction is zero.
- The `vpternlogd` computes any ternary Boolean function (e.g., A OR B OR C). The function is selected by a programmer by passing an immediate 8-bit integer value to the instruction. This instruction can replace several logical instructions, thus decreasing the instruction count.
- The multiply-add `vpaddubsw` instruction works over pairs of registers which are considered as arrays of thirty-two 16-bit block. Each 16-bit block is treated as a vector of two 8-bit integer values. Values from one register are treated as signed bytes $[-128, +127)$ and values from the other register as unsigned bytes $[0, 256)$.

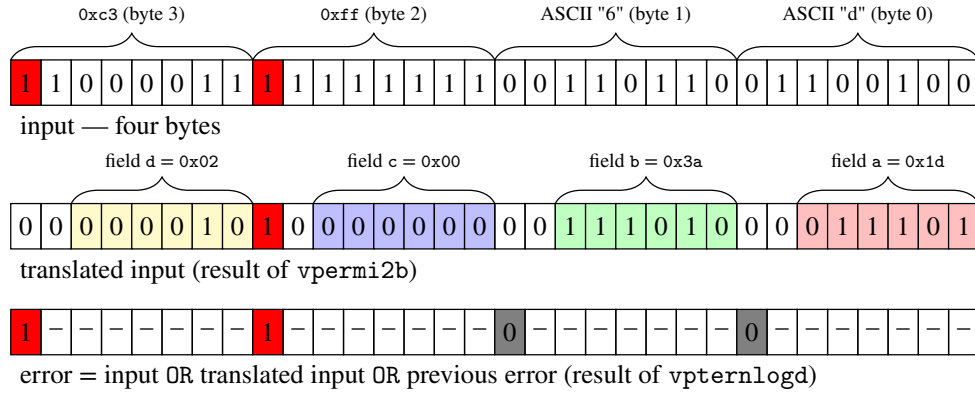


FIGURE 2 Decoding from ASCII into 6-bit values with error detection

We compute the scalar product of aligned pairs of 16-bit blocks: if the 8-bit integers x_1 and x_2 come from the first register, and the 8-bit integers y_1 and y_2 come from the second register, we output $x_1 \times y_1 + x_2 \times y_2$ as a 16-bit value.

- The `vpmaddwd` instruction works similarly to the `vpmaddubsw` instructions but uses 32-bit blocks and multiplying 16-bit signed integers instead.

We loop over 512-bit inputs, i.e. 64 ASCII characters. We first translate the ASCII characters into 6-bit values stored in separate bytes. At this stage, we must detect errors; that is, we must detect invalid ASCII characters outside of the base64 table. To translate the ASCII characters to 6-bit values while detecting bad characters, the `vpermi2b` instruction is ideally suited. We set all byte values in the two lookup arrays to byte values 0x80 except for values at indexes corresponding to base64 ASCII code points: 0x41, 0x42, ... which are set to their corresponding 6-bit value, as per Table 1. If all input values are allowable base64 ASCII characters, we get the correct 6-bit values stored in a corresponding byte. If any unallowed ASCII character is encountered, then it will result in the 0x80 byte value. Furthermore, any non-ASCII code point is characterized as an input byte with the most significant bit set. Thus if we take the resulting register from the `vpermi2b` instruction and that we compute the bitwise OR with the original input, the input contains an incorrect character if and only if we can detect a byte value with its most significant bit set. If we want to report errors immediately, we can make this check with the `vpmovb2m` instruction followed by a conditional branch instruction. However, we may postpone error reporting to the end, avoiding branching in the hot loop. We initialize a 512-bit ERROR register with zeros, and then repeatedly replace its value with its bitwise OR with the bitwise OR of the input register and of the result `vpermi2b` instruction. The two bitwise OR operations can be expressed with a single instruction `vpternlogd`. That is, the `vpmovb2m` instruction followed by a branch is replaced by a single instruction (`vpternlogd`). At the end of the processing, we can check the most significant bit values of the bytes of the ERROR register are zero. See Fig. 2.

The final stage of decoding consists in packing all 6-bit fields into a continuous array of 48 bytes. We proceed in two steps.

- We first need to pack four 6-bit values into 24-bit words. Thus we need to modify, in parallel, the sixteen 32-bit words contained in our working register. We achieve this result by two multiply-add instructions, operating over 32-bit words. We can represent our initial stage within each 32-bit words as follows:

[00ddddd | 00cccc | 00bbbbbb | 00aaaaaa] .

If we represent the four byte values as D, C, B, A , we want to first replace the sequence of two 8-bit values D, C , by the 16-bit value $D + C \times 2^6$; and similarly for the sequence B, A . We can get this result with the `vpmaddubsw` by combining our working register with a constant register made of the values $(0, 2^6, 0, 2^6, \dots)$. The result on a 32-bit word basis is

[0000cccc | cddddd | 0000aaaa | aabbbbbb] .

Finally, we need to combine successive 16-bit words, which we can do in a similar manner, this time using the `vpmaddwd` instruction and multiplying our working register with the constant $(2^{12}, 1, 2^{12}, 1, \dots)$. We get

[00000000 | aaaaabb | bbbcccc | cddddd] .

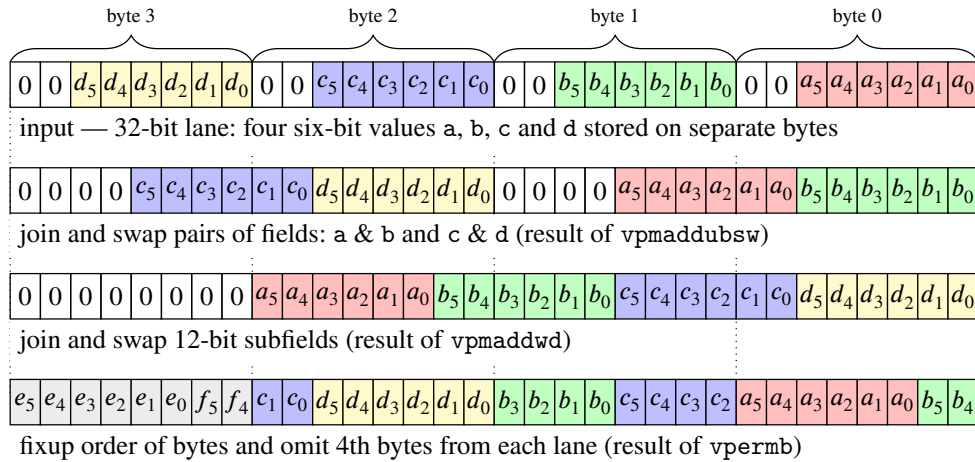


FIGURE 3 Vectorized decoding of three bytes

- One byte out of each 4 bytes is zero. We need to *pack* the result into 48 consecutive bytes, in the order specified by the base64 standard. We can achieve this result with the `vperm b` instruction, using the 48 indexes

6, 0, 1, 2, 9, 10, 4, 5, 12, 13, 14, 8, 22, 16, 17, 18, 25, 26, 20, 21, 28, 29, 30, 24, 38, 32, 33,
34, 41, 42, 36, 37, 44, 45, 46, 40, 54, 48, 49, 50, 57, 58, 52, 53, 60, 61, 62, 56.

There are 16 trailing bytes that are unused, so the last 16 indexes are irrelevant.

Thus the second stage consists in only three instructions: `vpmaddubsw`, `vpmaddwd`, and `vperm b`. See Fig. 3.

As with the encoding algorithm, we need handle the case where the input is not a multiple of 64 bytes with a final and separate code path.

4 | EXPERIMENTS

We summarize the characteristics of our hardware platform in Table 2. This processor has 32 kB of fast (L1) cache per core, 256 kB of secondary (L2) cache per core and 4 MB of last-level cache (L3). The system has a maximal copy (`memcpy`) bandwidth of ≈ 20 GB/s. On some Intel processors, AVX-512 instructions¹² are subject to downclocking: whenever AVX-512 instructions are executed, the processor lowers its frequency. We do not observe any downclocking on our processor, nor did we find any Intel documentation referencing downclocking for this processor. To test for downclocking, we used the `avx-turbo`¹³ benchmarking tool: it reports the processor frequency after issuing long sequences of expensive instructions of various types, including AVX-512 multiplication and floating-point instructions.

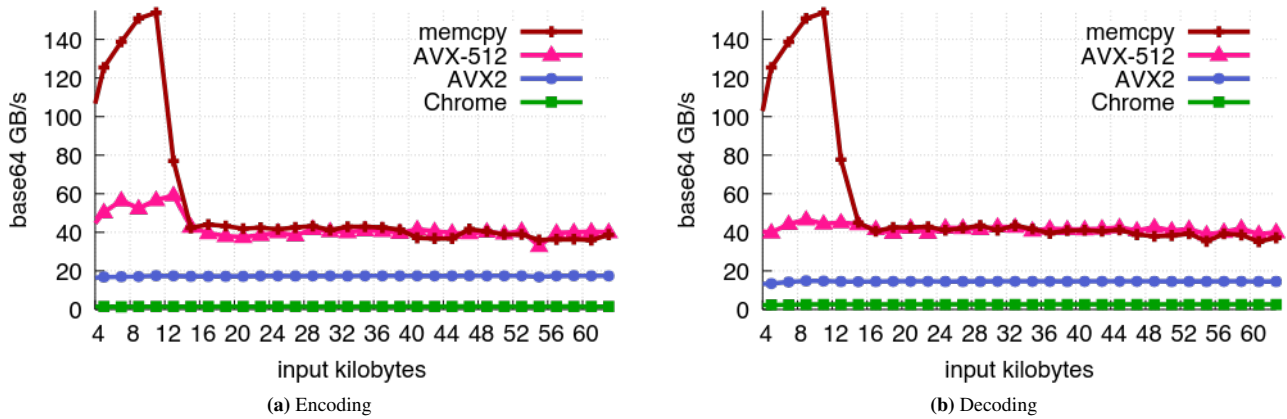
We implemented our software in C. To ease reproducibility, we make it freely available.¹ We compile our code using the GNU GCC 9.1 compiler with the `-O3 -march=cannonlake` flags. As a state-of-the-art conventional base64 codec, we use the library used by the Chrome browser. For comparison, we use an optimized AVX2 codec⁴ from our previous work. To measure the speed, we take 10 measures, compute the median time. Our timings include some fixed overhead costs such as the function call.

In Fig. 4, we present the encoding and decoding speed using random binary data as data source, varying the size from 1 kB to 64 kB. Our processor has 32 kB of L1 cache per core, so we expect that it is able to do a memory copy (`memcpy`) entirely in L1 cache as long as the input fits in 16 kB. We see that for inputs of about 10 kB, we reach a top copy speed of over 150 GB/s. This speed is slightly penalized by overhead (function calls and time measurements). The speed of the new AVX-512 coded is more than twice that of the state-of-the-art AVX2⁴ codec. This is especially apparent when the data fits in L1 cache (e.g., when the input is less or equal to 12 kB). The speed of the AVX-512 codec is limited to 40 GB/s for inputs larger than 16 kB—the same

¹<https://github.com/WojciechMula/base64-avx512>

TABLE 2 Hardware

Processor	Base Frequency	Max. Frequency	L1 data cache per core	Microarchitecture	Compiler
Intel i3-8121U	2.2 GHz	3.2 GHz	32 kB	Cannon Lake (2018)	GCC 9.1

**FIGURE 4** Speed in GB/s to encode or decode base64. Data volume is measured in base64 bytes. Speed is lower on tiny inputs due to fixed overheads.**TABLE 3** Decoding performance in GB/s

Source	bytes	memcpy	Chrome	AVX2	AVX-512
lena [jpg]	141 020	25	2.6	14	32
mandril [jpg]	247 222	18	2.6	14	25
Google logo [png]	2357	44	2.6	14	42
large [zip]	34 904 444	9.5	2.6	8.3	9.5

speed also limits the memory copy. For larger input, the speed of the AVX2 codec is 17 GB/s for encoding about 15.5 GB/s for decoding. The conventional codec (Chrome) encodes at 1.5 GB/s and decodes at 2.6 GB/s.

We do not expect the vectorized codecs (AVX2 and AVX-512) to be sensitive to the content of the input, keeping the size constant. Nevertheless we run benchmarks on standard images from image processing (Lena, mandril), a Google logo found to be base64 encoded in the Google search page and a large zip file. Our test data is available along with our software. See Table 3 for decoding performance. As expected, the results are similar as with random data. The decoder used in the Chrome browser has a constant speed of 2.6 GB/s, irrespective of the input size. For the Google logo which fits in L1 cache, the memory copy is faster, but otherwise our AVX-512 is even slightly faster than a memory copy. This is possible because we use the base64 size as a reference when computing the speed: to decode 1 GB of base64 data, we only need to write about 786 MB whereas a memory copy must still write a full 1 GB. For the large zip file, the AVX-512 is as fast as a memory copy: the benefit of the vectorized decoders (AVX2 and AVX-512) is less important in this case because we are limited by memory access. In practice, it might be preferable to process large files in small parts that fit in cache when possible to avoid having to write to RAM.

5 | CONCLUSION

A straight-forward extension of the best base64 codec using 256-bit registers (AVX2) to 512-bit registers (AVX-512) should reduce the instruction count by a factor of two in the best scenario. Yet, not counting load and store instructions, we reduce the

instruction count by a factor of seven (for decoding) and five (for encoding). The net result is that we can encode and decode base64 data at a speed of 40 GB/s when the data is in secondary (L2) cache. For many applications, we expect that encoding and decoding base64 might become a negligible computational burden with our AVX-512 approach: our speed is limited by data access. Our results illustrate the value of powerful SIMD instructions: our codec is 10 to 20 times faster than a highly optimized conventional codec.

The AVX-512 instructions are currently specific to Intel processors. However, wider and more powerful SIMD instructions are already available on some ARM processors¹⁴ via the SVE extension, with an even more powerful extension (SVE2) having been announced. We expect that such new instruction sets should be applicable to base64 decoding and encoding. Future work could also integrate fast base64 decoders inside vectorized parsers such as `simdjson`¹⁵.

ACKNOWLEDGMENTS

The work is supported by the Natural Sciences and Engineering Research Council of Canada under grant RGPIN-2017-03910.

References

1. Freed N, Borenstein NS. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. <https://tools.ietf.org/html/rfc2045>; 1996. Internet Engineering Task Force, Request for Comments: 2045.
2. Crane M, Lin J. An Exploration of Serverless Architectures for Information Retrieval. In: ICTIR '17. ACM; 2017; New York, NY, USA: 241–244
3. Cebrian JM, Natvig L, Jahre M. Scalability analysis of AVX-512 extensions. 2019. to appear in The Journal of Supercomputing.
4. Muła W, Lemire D. Faster Base64 Encoding and Decoding Using AVX2 Instructions. *ACM Trans. Web* 2018; 12(3): 20:1–20:26. doi: 10.1145/3132709
5. Josefsson S. The Base16, Base32, and Base64 Data Encodings. <https://tools.ietf.org/html/rfc4648>; 2006. Internet Engineering Task Force, Request for Comments: 4648.
6. Reinders J. AVX-512 Instructions. <https://software.intel.com/en-us/articles/intel-avx-512-instructions>; 2013.
7. Rucci E, Garcia Sanchez C, Botella Juan G, Giusti AD, Naiouf M, Prieto-Matias M. SWIMM 2.0: Enhanced Smith–Waterman on Intel’s Multicore and Manycore Architectures Based on AVX-512 Vector Extensions. *International Journal of Parallel Programming* 2019; 47(2): 296–316. doi: 10.1007/s10766-018-0585-7
8. Kataoka H, Yamashita K, Nakano K, Ito Y, Kasagi A, Tabaru T. An Efficient Convolutional Neural Network Computation using AVX-512 Instructions. *Bulletin of Networking, Computing, Systems, and Software* 2019; 8(2).
9. Lasch R, Oukid I, Dementiev R, May N, Demirsoy SS, Sattler KU. Fast & Strong: The Case of Compressed String Dictionaries on Modern CPUs. In: DaMoN’19. ACM; 2019; New York, NY, USA: 4:1–4:10
10. Zarubin M, Damme P, Kissinger T, Habich D, Lehner W, Willhalm T. Integer Compression in NVRAM-centric Data Stores: Comparative Experimental Analysis to DRAM. In: DaMoN’19. ACM; 2019; New York, NY, USA: 11:1–11:11
11. Kersten T, Leis V, Kemper A, Neumann T, Pavlo A, Boncz P. Everything You Always Wanted to Know About Compiled and Vectorized Queries But Were Afraid to Ask. *Proc. VLDB Endow.* 2018; 11: 2209–2222.
12. Intel 64 and IA-32 Architectures Optimization Reference Manual. Tech. Rep. 248966-033, Intel Corporation; 2016.
13. Downs T. avx-turbo: Test the non-AVX, AVX2 and AVX-512 speeds across various active core counts. <https://github.com/travisdowns/avx-turbo>; 2019.

14. Stephens N, Biles S, Boettcher M, et al. The ARM Scalable Vector Extension. *IEEE Micro* 2017; 37(2): 26-39. doi: 10.1109/MM.2017.35
15. Langdale G, Lemire D. Parsing Gigabytes of JSON per Second. *The VLDB Journal* (to appear).

