

High-Performance Decimal Floating-Point Arithmetic: Algorithms and Implementation in C++

MATTHEW BORLAND and CHRISTOPHER KORMANYOS

This article presents a comprehensive, portable C++ implementation of decimal floating-point arithmetic conforming to IEEE 754-2019 standards[8]. Our system offers three IEEE 754-compliant types, and three additional types that prioritize performance over strict standard adherence, providing flexible options for various computational needs. We describe in detail the Decimal system architecture, its standard library, and usage guidelines. The implementation incorporates novel algorithms for key operations, significantly improving performance in common use cases. Rigorous testing results demonstrate the system's correctness and IEEE 754 compliance. Performance benchmarks show competitive or superior results compared to existing implementations. This work addresses the growing demand for precise decimal arithmetic in financial, scientific, and engineering applications, offering a robust, efficient solution for C++ developers.

CCS Concepts: • General and reference → Design; Performance; • Mathematics of computing → Mathematical software performance; • Software and its engineering → Software architectures.

Additional Key Words and Phrases: C++, Decimal Floating Point, Object Oriented, Special Functions, System Architecture

ACM Reference Format:

Matthew Borland and Christopher Kormanyos. 2018. High-Performance Decimal Floating-Point Arithmetic: Algorithms and Implementation in C++. *ACM Trans. Math. Softw.* 37, 4, Article 111 (August 2018), 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Decimal floating-point arithmetic is crucial for many applications[4], particularly in financial and scientific computing. While several C++ decimal floating-point packages exist, they often lack IEEE 754[8] conformance, interoperability with the C++ Standard Template Library (STL), or both, and have limited portability. This paper presents a novel decimal system that addresses these limitations and advances decimal floating-point technology. Our system is standalone, relies on a minimal subset of the C++ STL, and has been tested on a wide range of devices, from S390X mainframes to AVR boards. It provides seamless interoperability with existing C++ standard types and full IEEE 754 conformance, features not collectively offered by any known package. This work significantly enhances the toolset available for high-precision decimal computations across diverse computing environments.

2 The Decimal System

2.1 Background

Decimal floating point is a method of representing floating point numbers using base-10 instead of base-2 like most are familiar with. Decimal floating point is not a new concept per se. Mechanical

Authors' Contact Information: Matthew Borland, matt@mattborland.com; Christopher Kormanyos, e_float@yahoo.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 1557-7295/2018/8-ART111
<https://doi.org/XXXXXXX.XXXXXXX>

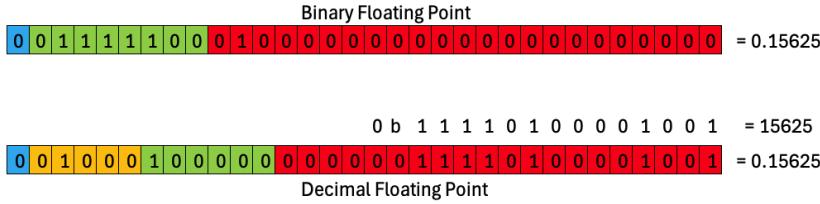


Fig. 1. Bit layouts of Binary and Decimal Floating Point Numbers

devices such as the abacus and slide rule use base-10 numbers. Early computers such as the IBM 650 used decimal floating point numbers prior to IEEE standardization[2]. Some modern architectures support decimal floating point in hardware such as IBM System Z[15]. The advantage of decimal floating point is that it can represent human readable floating point numbers exactly unlike binary floating point.

The major difference between the layout of binary and decimal floating-point numbers lies in their internal structure and how they represent the same numerical value. Figure 1 illustrates this difference by showing the bit layout of the same number in both formats.

Binary floating-point numbers consist of three parts:

- The sign bit (shown in blue)
- The exponent (shown in green)
- The significand (shown in red)

Decimal floating-point numbers, on the other hand, are composed of four distinct parts:

- The sign bit (shown in blue)
- The combination field (shown in orange)
- The exponent continuation (shown in green)
- The coefficient continuation (shown in red)

As depicted in Figure 1, these different structures allow the same numerical value to be represented in two fundamentally different ways. The binary format uses a base-2 system, while the decimal format uses a base-10 system, which affects how precisely certain decimal values can be represented and how rounding errors propagate in calculations. Due to these differences decimal can represent numbers exactly that binary cannot, for example the number 0.3 which with a C++ double would actually be: 0.3000000000000004. There are actually two different bit layouts for decimal floating point types: Binary Integer Significand Field (BID), and Densely Packed Decimal Significand Field (DPD)[8]. The above example is depicted in the DPD format which is typically for hardware implementation. For the remainder of this paper we will focus on the BID format as it is more natural for software implementations like ours.

2.1.1 Cohorts. Another idiosyncrasy that is found in decimal floating point types are called cohorts[8]. A cohort is all the different ways to represent the same value. For example one can represent the number 10 as: 0.1×10^2 , 1×10^1 , 10×10^0 , 100×10^{-1} , and so on. The ramification of this is we must normalize the significand and exponents of a number before we can attempt doing comparisons or other basic mathematical operations.

2.2 Provided Types

The decimal system provides 6 total types: `decimal32_t`, `decimal64_t`, `decimal128_t`, `decimal_fast32_t`, `decimal_fast64_t`, and `decimal_fast128_t`. The first three types are conformant to IEEE-754

standards while the latter three provided identical numerical results without the constraints of IEEE-754.

The three conformant types is the design specifications provided in IEEE-754 namely their properties:

Parameter	decimal32	decimal64	decimal128
Storage Width	32	64	128
Precision (decimal digits)	7	16	34
Max Exponent	96	384	6144
Exponent Bias	101	398	6176
Sign Width	1	1	1
Combination Field Width	11	13	17
Significand Continuation Width	20	50	110

Table 1. Decimal Floating-Point Format Parameters

The three so-called fast types (i.e. `decimal_fastXX_t`) have the same values of precision, range, and exponent as their analogous conformant type, but require more space. This is a similar approach to `std::uint_fastXX_t` types which have at least the same values as the standard types.

2.3 System Architecture

The decimal system architecture is robust and flexible. Each type is implemented completely independently of each other, but they share many of the same function implementations from the STL.

2.4 Using the System

Every effort was made during design and implementation to ensure that using the decimal system was straightforward and intuitive. The library contains zero dependencies, and uses only a small subsection of the C++ STL. The required language standard for the library is C++14.

```

1 #include <boost/decimal.hpp>
2 #include <iostream>
3 #include <iomanip>
4
5 int main()
6 {
7     using namespace boost::decimal;
8
9     constexpr decimal32 val_1 {100};           // Construction from an
10    integer
11    constexpr decimal32 val_2 {10, 1};          // Construction from an
12    integer and exponent
13    constexpr decimal32 val_3 {1U, 2, false}; // Construction from an
14    integer, exponent, and sign
15
16    std::cout << "Val_1: " << val_1 << '\n'
17        << "Val_2: " << val_2 << '\n'
18        << "Val_3: " << val_3 << '\n';

```

```

16
17     if (val_1 == val_2 && val_2 == val_3 && val_1 == val_3)
18     {
19         std::cout << "All equal values" << std::endl;
20     }
21
22     constexpr decimal64 val_4 {decimal64{2, -1} + decimal64{1, -1}};
23     constexpr double float_val_4 {0.2 + 0.1};
24     const decimal64 val_5 {float_val_4}; // Explicit Conversion from
25     double
26
27     std::cout << std::setprecision(17) << "Val_4: " << val_4 << '\n'
28             << "Float: " << float_val_4 << '\n'
29             << "Val_5: " << val_5 << '\n';
30
31     if (val_4 == val_5)
32     {
33         std::cout << "Floats are equal" << std::endl;
34     }
35     else
36     {
37         std::cout << "Floats are not equal" << std::endl;
38     }
39
40     return 0;

```

Listing 1. Basic Usage

In Listing 1 we show how to construct the type and that it works just like a built-in floating point type. The following is an example that leverages one of decimal-floating points main strengths and one that we expect many people will use.

```

1 #include <boost/decimal.hpp>
2 #include <iostream>
3 #include <cassert>
4
5 int main()
6 {
7     using namespace boost::decimal;
8
9     decimal64 val {0.25}; // Construction from a double (not recommended
10    but explicit construction is allowed)
11
12    char buffer[256];
13    auto r_to = to_chars(buffer, buffer + sizeof(buffer) - 1, val);
14    assert(r_to); // checks std::errc()
15    *r_to.ptr = '\0';
16
17    decimal64 return_value;

```

```

17     auto r_from = from_chars(buffer, buffer + std::strlen(buffer),
18     return_value);
19     assert(r_from);
20
21     assert(val == return_value);
22
23     std::cout << " Initial Value: " << val << '\n'
24             << "Returned Value: " << return_value << std::endl;
25
26     return 0;
27 }
```

Listing 2. Basic Usage

In Listing 2 we show how to serialize and parse numbers. These techniques and functions are an extension of Boost.Charconv[3].

3 Implementation Details

3.1 IEEE 754 Conformant Decimal Types

The decimal system provides three IEEE-754 compliant types: decimal32, decimal64, and decimal128. Each of these are constructed using a single unsigned integer of the same width to hold the bits. For example decimal32 consists of a single `std::uint32_t`. decimal128 uses a custom implementation of a 128-bit unsigned integer for portability reasons rather than relying on the existence of `unsigned __int128`. The discussion of implementing big integers is outside the scope of the paper, but papers and implementations can be found in numerous places[3][12][10].

The decoding step of decimal32 is somewhat involved because the values of the significand and exponent depend on the bits in the combination field. Our first case is where the bits in the combination field are in the form 00XXX, 01XXX, and 10XXX.

```

1 //      Comb.   Exponent          Significand
2 // s       eeeeeeee    tttttttttttttttttt
3 // s   11   eeeeeeee  [100]ttttttttttttttttt
```

Listing 3. Combination Field Pattern 1

In Listing 3 we can see that with these combination field bit patterns that most of the time a decimal floating point number is actually in the same bit layout as a binary floating point number. The special case is when we need more than 23 bits to represent the significand of our number. In this case we use 2 bits in the combination field to imply 3 bits in the significand which gives us the full range of significand values.

The final case is where the bits in the combination field are used for non-finite numbers:

```

1 static constexpr std::uint32_t d32_inf_mask  = UINT32_C(0x78000000);
2 static constexpr std::uint32_t d32_nan_mask  = UINT32_C(0x7C000000);
3 static constexpr std::uint32_t d32_snan_mask = UINT32_C(0x7E000000);
```

Listing 4. Combination Field Pattern 3

Since this combination field pattern is only for non-finite numbers it makes the implementation of `isnan`, `isinf`, etc. straightforward. We can also see the signbit remains unused so we can represent signed infinities. One can also use the remaining exponent and significand bits to provide a payload to NaN analogous to the use of `std::nan`[6].

Now that we have seen what each bit pattern of the combination field corresponds to; the full implementation of decoding the significand then becomes:

```

1 constexpr auto decimal32_t::full_significand() const noexcept ->
2     significand_type
3 {
4     significand_type significand {};
5
6     if ((bits_ & detail::d32_comb_11_mask) == detail::d32_comb_11_mask)
7     {
8         constexpr std::uint32_t implied_bit {UINT32_C(0
9             b1000000000000000000000000000000)};
10        significand = implied_bit | (bits_ & detail::
11            d32_11_significand_mask);
12    }
13    else
14    {
15        significand = bits_ & detail::d32_not_11_significand_mask;
16    }
17 }
```

Listing 5. Decoding decimal32_t significand

Encoding the value is a similar to the decoding process found in Listing 5 , but more complex process in that we need to work backwards through the steps to figure out what the value of the combination field needs to be, and then we use a series of masks to write the value.

3.2 IEEE 754 Fast Types

Now that we have discussed the three IEEE-754 conformant types we will turn our attention to much more performant, but non-compliant types. The library offers: decimal32_fast, decimal64_fast, and decimal128_fast. These types are similar to how instead of `std::uint32_t` you can use `std::uint_fast32_t` which on x86_64 platforms generally aliases to `std::uint64_t`. This is the classic optimization of trading space for time. Again we will focus on the 32-bit type for clarity and simplicity. Unlike decimal32 consisting of a data `std::uint32_t`, decimal32_fast actually contains its internal state in a structure.

```

1 class decimal32_fast final
2 {
3 public:
4     using significand_type = std::uint_fast32_t;
5     using exponent_type = std::uint_fast8_t;
6     using biased_exponent_type = std::int_fast32_t;
7
8 private:
9     // In regular decimal32 we have to decode the 24 bits of the
10    significand and the 8 bits of the exp
```

```

10 // Here we just use them directly at the cost of at least 2 extra
11 // bytes of internal state
12 // since the fast integer types will be at least 32 and 8 bits
13 // respectively
14
15 significand_type significand_{};
16 exponent_type exponent_{};
17 bool sign_{};
18
19 // Continued
20
21 };

```

Listing 6. Internal state of decimal32_fast

In Listing 6 we see that now instead of having to encode and decode the number every time we perform an operation we can now just access the value directly:

```

1 constexpr auto full_significand() const noexcept -> significand_type
2 {
3     return significand_;
4 }

```

Listing 7. Decoding decimal32_fast significand

Now compare the decoding process found in the conformant type in Listing 5 to that in Listing 7. Encoding the value is similarly trivial. Each of the fast types use the same approach, but with different types used for the internal state to ensure that the range and precision is equal to that of the IEEE 754 conformant types.

An additional optimization for the fast types is that exponent and significand are stored in normalized form. As discussed in Section 2.1.1 we would have to normalize the values of the exponent and significand prior to any comparison or mathematical operation. Rather than having to normalize each time we perform an operation we do it exactly one time, when the number is constructed. This is beneficial because normalization has time complexity of $O(\log(n))$ whereas most basic operations are only $O(1)$.

```

1 // Converts the significand to full precision to remove the effects of
2 // cohorts
3 template <typename TargetDecimalType, typename T1, typename T2>
4 constexpr auto normalize(T1& significand, T2& exp, bool sign = false)
5     noexcept -> void
6 {
7     constexpr auto target_precision {detail::precision_v<
8         TargetDecimalType>};
9     const auto digits {num_digits(significand)};
10
11    if (digits < target_precision)
12    {
13        const auto zeros_needed {target_precision - digits};
14        significand *= pow10(static_cast<T1>(zeros_needed));
15        exp -= zeros_needed;
16    }
17
18    else if (digits > target_precision)
19    {
20        const auto zeros_needed {target_precision - digits};
21        significand /= pow10(static_cast<T1>(zeros_needed));
22        exp += zeros_needed;
23    }
24 }

```

```

15  {
16      const auto excess_digits {digits - (target_precision + 1)};
17      significand /= pow10(static_cast<T1>(excess_digits));
18      // Perform final rounding according to the fenv rounding mode
19      exp += detail::fenv_round<TargetDecimalType>(significand, sign ||
20          significand < 0U) + excess_digits;
21 }

```

Listing 8. Normalization Method

As we can see in the above listing in the event that our significand has too few digits we append a fixed number of zeros, and if it's too many (e.g. in the constructor) we must remove excess and round appropriately.

3.3 Basic Operations

Now that we have discussed how to encode and decode the sign, exponent, and significand of a decimal type number we will cover basic operations.

3.3.1 Comparisons. All types provided by this system support $>$, $>=$, $!=$, $==$, $<=$, $<$, and when using C++20 $<=>$. As discussed in Section 2.1.1 we will need to normalize the values of the exponent and significand to ensure fair comparisons. Once we have normalized our values using the methods from Listing 8 we are then able to implement these operations by comparing the values of the significand, exponent, and sign. We must also check for non-finite values such as NaN which aren't comparable values.

3.3.2 Add and Sub. We are able to utilize the property that $a + b == a - (-b)$, as well as commutativity of operations ($a + b == b + a$) to unify our implementations of addition and subtraction.

3.3.3 Mul.

3.3.4 Div.

3.4 Standard Library

3.4.1 Special Functions. The implementation of transcendental, and C++17 mathematical special functions[5][9] extensively uses Páde Approximations and Remez Polynomials.

4 Results

4.1 Testing and Precision

4.2 Performance

As this system is implemented in software it will never be as performant as binary floating point is on computing systems with floating-point hardware. For example the table below shows the runtime difference for comparison operations between `float`, `double`, and the types provided by this system:

The gap in performance increases even further for the multiplication operation which is typically only a few cycles in hardware[1]:

For floating-point operations that are implemented in software like parsing and serializing numbers the performance gap is quite smaller:

Type	Runtime (μ s)	Ratio to double
float	8,587	1.376
double	6,240	1.000
decimal32	275,597	44.166
decimal64	296,929	47.587
decimal32_fast	99,664	15.972
decimal64_fast	102,132	16.367

Table 2. Comparison of runtime and ratio for comparison operations

Type	Runtime (μ s)	Ratio to double
float	1,646	0.957
double	1,720	1.000
decimal32	313,219	182.104
decimal64	583,818	339.429
decimal32_fast	86,093	50.054
decimal64_fast	333,582	193.943

Table 3. Comparison of runtime and ratio for multiplication operations

Type	Runtime (μ s)	Ratio to double
float	235,816	0.953
double	247,307	1.000
decimal32	366,682	1.483
decimal64	485,965	1.965

Table 4. Comparison of runtime and ratio for from_chars

Type	Runtime (μ s)	Ratio to double
float	316,300	1.040
double	304,272	1.000
decimal32	406,053	1.335
decimal64	678,451	2.230

Table 5. Comparison of runtime and ratio for to_chars

5 Conclusion and Outlook

The portable C++ decimal system has been presented. The system allows for users to easily employ decimal-floating point numbers in their existing code bases. For the first time a complete and interoperable system has been fully provided.

During the course of our research and implementation we have limited ourselves to 32, 64, and 128-bit types. Further research can be conducted into making higher precision types as the properties of such numbers are specified in IEEE 754. We could also expand our range of compatibility to allow the types to be massively parallelized such as providing support for CUDA devices.

Acknowledgments

To the C++ Alliance for sponsoring the development of this library.

References

- [1] ARM Limited. 2023. *ARM Architecture Reference Manual*. ARM Limited. <https://developer.arm.com/documentation/ddi0487/latest> ARM DDI 0487J.a (ID050623).
- [2] Nelson H. F. Beebe. 2017. Historical floating-point architectures. In *The Mathematical-Function Computation Handbook - Programming Using the MathCW Portable Software Library* (1 ed.). Springer International Publishing AG, Salt Lake City, UT, USA, Chapter H, 948. <https://doi.org/10.1007/978-3-319-64110-2>
- [3] Matt Borland, Peter Dimov, Junekey Jeon, Alexander Grund, Andrzej Krzemieński, Dmitry, Vinnie Falco, and Sam Darwin. 2024. *boostorg/charconv: Boost 1.86.0*. <https://doi.org/10.5281/zenodo.13323694>
- [4] Michael F. Cowlishaw. 2003. Decimal Floating-Point: Algorithm for Computers. In *Proceedings of the 16th IEEE Symposium on Computer Arithmetic*. IEEE, 104–111. <https://doi.org/10.1109/ARITH.2003.1207670>
- [5] cppreference.com contributors. 2024. Special mathematical functions. https://en.cppreference.com/w/cpp/numeric/special_functions Accessed: 2024-08-23.
- [6] cppreference.com contributors. 2024. std::nan, std::nanf, std::nanl. <https://en.cppreference.com/w/cpp/numeric/math/nan> Accessed: 2024-08-23.
- [7] John F. Hart, E. W. Cheney, Charles L. Lawson, Hans J. Maehly, Charles K. Mesztenyi, John R. Rice, Henry G. Thacher, and Christoph Witzgall. 1968. *Computer Approximations*. John Wiley & Sons, New York.
- [8] IEEE. 2019. *IEEE Standard for Floating-Point Arithmetic*. Standard IEEE 754-2019. IEEE Computer Society, New York, NY, USA. <https://doi.org/10.1109/IEEEESTD.2019.8766229>
- [9] ISO/IEC. 2023. *Programming Languages – C++*. Standard ISO/IEC 14882:2023. International Organization for Standardization, Geneva, Switzerland. Sixth edition.
- [10] Donald E. Knuth. 1997. *The Art of Computer Programming* (3rd ed.). Vol. 1-4B. Addison-Wesley, Reading, MA.
- [11] Christopher Kormanyos. 2011. Algorithm 910: A Portable C++ Multiple-Precision System for Special-Function Calculations. *ACM Transactions on Mathematical Software (TOMS)* 37, 4, Article 45 (feb 2011), 27 pages. <https://doi.org/10.1145/1916461.1916469>
- [12] Daniel Lemire. 2021. Number Parsing at a Gigabyte per Second. *Software: Practice and Experience* 51, 8 (2021), 1766–1785. <https://doi.org/10.1002/spe.2914>
- [13] Jean-Michel Muller. 2016. *Elementary Functions: Algorithms and Implementation* (3 ed.). Birkhäuser, Boston. <https://doi.org/10.1007/978-1-4899-7983-4>
- [14] Jean-Michel Muller, Nicolas Brunie, Florent de Dinechin, Claude-Pierre Jeannerod, Mioara Joldes, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, and Serge Torres. 2018. *Handbook of Floating-Point Arithmetic* (2 ed.). Birkhäuser, Cham, Switzerland. <https://doi.org/10.1007/978-3-319-76526-6>
- [15] Eric M Schwarz, John M Kapernick, and Mike F Cowlishaw. 2009. Decimal floating-point support on the IBM System z10 processor. *IBM Journal of Research and Development* 53, 1 (2009), 4:1–4:10. <https://doi.org/10.1147/JRD.2009.5388557>

Received XX February XXXX; revised XX March XXXX; accepted X June XXXX