Radix Conversion for IEEE754-2008 Mixed Radix Floating-Point Arithmetic

Olga Kupriianova
UPMC Paris 6 – LIP6 – PEQUAN team
4, place Jussieu
75252 Paris Cedex 05, France
Email: olga.kupriianova@lip6.fr

Christoph Lauter
UPMC Paris 6 – LIP6 – PEQUAN team
4, place Jussieu
75252 Paris Cedex 05, France
Email: christoph.lauter@lip6.fr

Jean-Michel Muller
CNRS – ENS Lyon – Université de Lyon
46, allée d’Italie
69364 Lyon Cedex 07, France
Email: jean-michel.muller@ens-lyon.fr

Abstract—Conversion between binary and decimal floating-point representations is ubiquitous. Floating-point radix conversion means converting both the exponent and the mantissa. We develop an atomic operation for FP radix conversion with simple straight-line algorithm, suitable for hardware design. Exponent conversion is performed with a small multiplication and a lookup table. It yields the correct result without error. Mantissa conversion uses a few multiplications and a small lookup table that is shared amongst all types of conversions. The accuracy changes by adjusting the computing precision.

I. INTRODUCTION

Humans are used to operate decimals while almost all the hardware is binary. According to IEEE754-2008 norm [1] a floating point number is represented as $\beta^E \cdot m$, where $\beta^{p-1} \leq m \leq \beta^p - 1$, $p$ is precision, $m \in \mathbb{N}$ is mantissa, $E \in \mathbb{Z}$ is exponent and $\beta$, the base or radix, is either two or ten. When the base $\beta = 2$, we have binary floating point (FP) numbers, when $\beta = 10$, the decimal one. However, most of hardware is binary, so the decimal mantissas are actually coded in binary. The formats for both radices differ by the length of stored numbers. Standardization of decimal FP arithmetic brings new challenges, e.g. supporting decimal transcendental functions with essentially binary hardware [2]. In [2] in order to evaluate decimal transcendental function the format conversion is used twice. The IEEE standard requires [1] the implementation of all the operations for different formats, but only for the operands of the same radix. The format does not require any mixed radix operations, i.e. one of the operands is binary, the other is decimal. Mixed radix arithmetic is currently being developed, although there are already some approaches published [3], [4].

Floating point radix conversion (from binary to decimal and vice versa) is a widespread operation, the simplest examples are the scanf and printf functions. It could also exist as an operation for financial applications or as a “precomputing step” for mixed radix operations. The radix conversion is used in number conversion operations, and implicitly in scanf and printf operations.

The current implementations of scanf and printf are correct only for one rounding mode and allocate a lot of memory. In this paper we develop a unified atomic operation for the conversion, so all the computations can be done in integer with the precomputed memory consumption.

While radix conversion is a very common operation, it comes in different variants that are mostly coded in ad-hoc way in existing code. However, radix conversion always breaks down into to elementary steps: determining an exponent of the output radix and computing a mantissa in the output radix. Section II describes the 2-steps approach of the radix conversion, section III contains the algorithm for the exponent computation, section IV presents a novel approach of raising 5 to an integer power used in the second step of the radix-conversion that computes the mantissa. Section V contains accuracy bounds for the algorithm of raising five to a huge power, section VI describes some implementation tricks and presents experimental results.

II. TWO-STEPS RADIX CONVERSION ALGORITHM

Conversion from a binary FP representation $2^E \cdot m$, where $E$ is the binary exponent and $m$ is the mantissa, to a decimal representation $10^F \cdot n$, requires two steps: determination of the decimal exponent $F$ and computation of the mantissa $n$. The conversion back to binary is pretty similar except of an extra step that will be explained later. Here and after consider the normalized mantissas $n$ and $m$: $10^{p_1} \leq n \leq 10^{p_1} - 1$ and $2^{p_2} \leq m \leq 2^{p_2} - 1$, where $p_1$ and $p_2$ are the decimal and binary precisions respectively. The exponents $E$ and $F$ are bounded by some values depending on the IEEE754-2008 format.

In order to enclose the converted decimal mantissa $n$ into one decade, for a certain output precision $p_{10}$, the decimal exponent $F$ has to be computed [5] as follows:

$$F = \left\lfloor \log_{10}(2^E \cdot m) \right\rfloor - p_{10} + 1. \quad (1)$$

The most difficult thing here is the evaluation of the logarithm: as the function is transcendental, the result is always an approximation and function call is extremely expensive. Present algorithm computes the exponent [1] for a new-radix floating-point number only with a multiplication, binary shift, a precomputed constant and a lookup table (see section III).

Once $F$ is determined, the mantissa $n$ is given as

$$n = *_{p_{10}} \left( \frac{2^E \cdot m}{10^F} \right), \quad (2)$$

where $*_{p_{10}}$ corresponds to the current rounding mode (to the nearest, rounding down, or rounding up [1]). The conversions are always done with some error $\varepsilon$, so the following relation is
fulfilled: $10^F \cdot n = 2^E \cdot m \cdot (1 + \varepsilon)$. In order to design a unique algorithm for all the rounding modes it is useful to compute $n^*$, such that $10^F \cdot n^* = 2^E \cdot m$. Thus, we get the following expression for the decimal mantissa:

$$n^* = 2^E \cdot F^5 \cdot m$$

As $2^E - F$ is a simple binary shift and the multiplication by $m$ is small, the binary-to-decimal mantissa conversion reduces to compute the leading bits of $5^F$.

The proposed ideas apply with minor changes to decimal-to-binary conversion: the base of the logarithm is 2 on the exponent computation step and one additional step is needed: for the mantissa computation the power $5^F$ is required instead of $5^P$.

III. LOOP-LESS EXPONENT DETERMINATION

The current implementations of the logarithm function are expensive and produce approximated values. However, some earlier conversion approaches computed this approximation by Taylor series or using iterations. Here the exponent for the both conversions is computed exactly neither with any function call nor any polynomial approximation.

After performing one transformation step, the input can be rewritten as following:

$$F = \lfloor E \cdot \log_{10}(2) \rfloor + \lfloor \log_{10}(m) \rfloor + \lfloor \log_{10}(m) \rfloor - p_{10} + 1,$$

where $\{x\} = x - \lfloor x \rfloor$, the fractional part of the number $x$.

As the binary mantissa $m$ is normalized in one binade $2^{p-1} \leq m < 2^p$, we can assume that it lies entirely in one decade. If it is not the case, we can always scale it a little bit. The inclusion in one decade means that $\lfloor \log_{10}(m) \rfloor$ stays the same on the whole interval. So, for the given format one can precompute and store this value as a constant. Thus, it is possible to take the integer number $\lfloor \log_{10}(m) \rfloor$ out of the floor operation in the previous equation. After representing the first summand as a sum of its integer and fractional parts, we have the following expression under the floor operation:

$$\lfloor E \cdot \log_{10}(2) \rfloor + \{E \cdot \log_{10}(2)\} + \{\log_{10}(m)\}.$$

Here we add two fractional parts to an integer. We add something that is strictly less than two, so under the floor operation we have either an integer plus some small fraction that will be thrown away, or an integer plus one plus small fraction. Thus, we can take the fractional parts out of the floor brackets adding a correction $\gamma$:

$$[E \cdot \log_{10}(2)] + \gamma, \gamma \in \{0, 1\}.$$

This correction $\gamma$ equals to 1 when the sum of two fractional parts from the previous expression exceeds 1, or mathematically:

$$E \cdot \log_{10}(2) - [E \cdot \log_{10}(2)] + \log_{10}(m) - [\log_{10}(m)] \geq 1.$$

Due to the logarithm function the expression on the left is strictly monotonous (increasing). This means that we need only one threshold value $m^*(E)$, such that $\forall m \geq m^*(E)$ the correction $\gamma = 1$. As we know the range for the exponents $E$ beforehand, we can store the critical values $m^*(E) = 10^{\left(\lfloor E \cdot \log_{10}(2) - [E \cdot \log_{10}(2)] \rfloor + [\log_{10}(m)] \right)}$ in a table.

There is a technique proposed in [9] to compute $[E \cdot \log_{10}(2)]$ with a multiplication, binary shift and the use of a precomputed constant. So, finally the value of the decimal exponent can be obtained as

$$F = \lfloor E \cdot \log_{10}(2) \rfloor \cdot 2^L + \lfloor \log_{10}(m) \rfloor - p_{10} + 1 + \gamma \quad (3)$$

The algorithm pseudocode is provided below.

There is a technique proposed in [9] to compute $[E \cdot \log_{10}(2)]$ with a multiplication, binary shift and the use of a precomputed constant. So, finally the value of the decimal exponent can be obtained as

$$F = \lfloor E \cdot \log_{10}(2) \rfloor \cdot 2^L + \lfloor \log_{10}(m) \rfloor - p_{10} + 1 + \gamma \quad (3)$$

The algorithm pseudocode is provided below.

There is a technique proposed in [9] to compute $[E \cdot \log_{10}(2)]$ with a multiplication, binary shift and the use of a precomputed constant. So, finally the value of the decimal exponent can be obtained as

$$F = \lfloor E \cdot \log_{10}(2) \rfloor \cdot 2^L + \lfloor \log_{10}(m) \rfloor - p_{10} + 1 + \gamma \quad (3)$$

The algorithm pseudocode is provided below.
### Table I. TABLE SIZE FOR EXPONENT COMPUTATION STEP

<table>
<thead>
<tr>
<th>Format</th>
<th>Table size</th>
</tr>
</thead>
<tbody>
<tr>
<td>binary32</td>
<td>554 bytes</td>
</tr>
<tr>
<td>binary64</td>
<td>8392 bytes</td>
</tr>
<tr>
<td>binary128</td>
<td>263024 bytes</td>
</tr>
<tr>
<td>decimal32</td>
<td>792 bytes</td>
</tr>
<tr>
<td>decimal64</td>
<td>6294 bytes</td>
</tr>
<tr>
<td>decimal128</td>
<td>19713 bytes</td>
</tr>
</tbody>
</table>

IV. COMPUTING THE MANTISSA WITH THE RIGHT ACCURACY

As it was mentioned, the problem on the second step is the computation of the value $5^B$ with some bounded exponent $B \in \mathbb{N}$. If the initial range for the exponent of five contains negative values, we compute $5^{B+B}$, where $B$ is chosen in order to make the range for the exponents nonnegative. In this case we store the leading bits of $5^{-B}$ as a constant and after computing $5^{B+B}$ with the proposed algorithm, we multiply the result by the constant.

In this section we propose an algorithm for raising five to a huge natural power without rational arithmetic or divisions. The range for these natural exponents $B$ is determined by the input format, e.g. for the conversion from binary64 the range is about six hundred.

We propose to perform several Euclidean divisions in order to represent the number $B$ the following way:

$$B = 2^{n_k} \cdot q_k + 2^{n_{k-1}} q_{k-1} + \ldots + 2^{n_1} q_1 + q_0,$$  \hspace{1cm} (4)

where $0 \leq q_0 \leq 2^{n_1} - 1$, $n_k \geq n_{k-1}$, $k \geq 1$. The mentioned divisions are just a chain of binary shifts. All the quotients are in the same range and we assume that the range for $q_0$ is the largest one, so we have $q_i \in [0; 2^{n_1} - 1]$, $0 \leq i \leq k$. Once the exponent is represented as $B$, computation $5^B$ is done with the respect to the following expression:

$$5^B = (5q_k)^2^{n_k} \cdot (5q_{k-1})^2^{n_{k-1}} \cdot \ldots \cdot (5q_1)^2^{n_1} \cdot 5^{q_0}$$ \hspace{1cm} (5)

Let us analyze how the proposed formula can simplify the algorithm of raising five to the power $B$. We mentioned that all the quotients $q_i$ are bounded. By selecting the parameters $k$ and $n_i$, we can make these quotients small, so the values $5^{q_i}$ can be stored in a table. Then, each factor in (5) can be found on Fig. 1, the pseudocode for squarings is in algorithm 2 and for the final multiplication step in algorithm 3.

Algorithm 2: Squaring with shifting $\lambda$ last bits

```
1 m ← 1;
2 for i ← k to 1 do
3   m ← \((m \cdot v_i) \cdot 2^{-\lambda}\);
4 end
5 m ← \((m \cdot 5^{q_0}) \cdot 2^{-\lambda}\);
6 m ← m \cdot 2\((2^{n_k}-1) + (2^{n_{k-1}}-1) + \ldots + (2^{n_1}-1) + k) \cdot 5^{q_0} - \sum_{i=k}^{1} \sigma_i;
7 s ← \sum_{i=k}^{1} (n_i\left(\left\lfloor \log_2(5^{q_i}) \right\rfloor - p + 1\right) + \left\lfloor \log_2(5^{q_0}) \right\rfloor - p + 1;
8 result ← m \cdot 2^s;
```

Algorithm 3: Final multiplication step

```
1 \text{input: } n_j, v_j = 5^{q_j};
2 \sigma_j ← 0;
3 for i ← 1 to n_j do
4   v_j ← v_j^2 \cdot 2^{-\lambda};
5   shiftNeeded ← 1 - \lfloor v_j \cdot 2^\lambda \rfloor //get the first bit;
6   v_j ← v_j \ll shiftNeeded;
7   \sigma_j ← 2 \cdot \sigma_j + \text{shiftNeeded};
8 end
9 result ← v_j \cdot 2^{-\sigma_j} \cdot 2^{(2^n-1-\lambda)};
```

There is still one detail in algorithm 2 that was not explained: the correction $\sigma_j$. The mantissa of the input number is represented as a binary number bounded by one binade (for both, binary and decimal formats). Assume that we operate the conversions in the range $[2^{p-1}, 2^p)$. After each squaring we can get a value less than infimum of this range. So, if the first bit of the intermediate result after some squarings is 0, we shift it to the left.

The described algorithm is applied $k$ times to each factors in (5). Then the last step is to multiply all the factors starting from the largest power like in listing below.

The whole algorithm schema is presented on Fig. 1. Depending on the range of $B$ one can represent it in different manner, but for our conversion tasks the ranges for $B$ were not that large, so the numbers $n_j$ were not more than 10 and the loops for squarings can be easily unrolled. For instance, for the conversions from binary32, binary64, decimal32 and decimal64 one can use the expansion of $B$ of the following form:

$$B = 2^8 \cdot q_2 + 2^4 \cdot q_1 + q_0$$

V. ERROR ANALYSIS

In order to compute the mantissa we use integer arithmetic but on each squaring/multiplication step we throw away a certain quantity of bits. So the final error is due to these right shiftings on each multiplication step.

We have errors only due to the multiplications, and as we do a lot of them, we need to define $N$ as the number of all the multiplications (squaring is just a particular case of multiplication). For each $i$-th factor ($1 \leq i \leq N$) in (5) we need to perform $n_i$ squarings, thus it gives us $n_i$ multiplications.
Thus, the relative error of the computations is
\[ N = \sum_{i=1}^{k} n_i + k. \]

So, the result is a product of \( N \) factors and on each step we have some relative error \( \varepsilon_i \). This means, that if we define \( y \) as the exact product without errors, then what we really compute in our algorithm can be represented as following:

\[ \hat{y} = y \prod_{i=1}^{N} (1 + \varepsilon_i). \]

Thus, the relative error of the computations is

\[ \varepsilon = \frac{\hat{y} - y}{y} = \prod_{i=1}^{N} (1 + \varepsilon_i) - 1. \]

Let us prove a lemma that will help us to find the bounds for the relative error of the result.

**Lemma 1.** Let \( N \geq 3 \), \( 0 \leq \varepsilon < 1 \) and \( |\varepsilon_i| \leq \varepsilon \) for all \( i \in [1, N] \). Then the following holds:

\[ \prod_{i=1}^{N} (1 + \varepsilon_i) - 1 \leq (1 + \varepsilon)^N - 1. \]

**Proof:** This inequality is equivalent to the following:

\[ -(1 + \varepsilon)^N + 1 \leq \prod_{i=1}^{N} (1 + \varepsilon_i) - 1 \leq (1 + \varepsilon)^N - 1. \]

The proof of the right side is trivial. From the lemma condition we have \( -\varepsilon \leq \varepsilon_i \leq \varepsilon \), which is the same as \( 1 - \varepsilon \leq 1 + \varepsilon_i \leq 1 + \varepsilon + 1 \) for arbitrary \( i \) from the interval \([1, N]\). Taking into account the borders for \( \varepsilon \), we get that \( 0 < (1 + \varepsilon_i) < 2 \) for all \( i \in [1, N] \). This means that we can multiply the inequalities \( 1 + \varepsilon_i \leq 1 + \varepsilon = 1 + \varepsilon_j \) with \( j \neq i \). After performing \( N - 1 \) such multiplications and taking into account that \( 1 + \varepsilon_i \leq \bar{\varepsilon} + 1 \), we get the following:

\[ \prod_{i=1}^{N} (\varepsilon_i + 1) \leq (\bar{\varepsilon} + 1)^N. \]

So, the right side is proved.

The same reasoning applies for the left bounds from the lemma condition, and the family of inequalities \( 1 - \bar{\varepsilon} \leq \varepsilon_i + 1 \) leads to the condition:

\[ (1 - \bar{\varepsilon})^N - 1 \leq \prod_{i=1}^{N} (1 + \varepsilon_i) - 1. \]

So, in order to prove the lemma we have to prove now that

\[ -(1 + \bar{\varepsilon})^N + 1 \leq (1 - \bar{\varepsilon})^N - 1. \]

After regrouping the summands we get the following expression to prove:

\[ 2 \leq (1 + \bar{\varepsilon})^N + (1 - \bar{\varepsilon})^N. \]

Using the binomial coefficients this transforms to

\[ 2 \leq 1 + \sum_{i=1}^{N} \binom{N}{i} \bar{\varepsilon}^i + 1 + \sum_{i=1}^{N} \binom{N}{i} (-\bar{\varepsilon})^i \]

On the right side of this inequality we always have the sum of 2 and some nonnegative terms. So, the lemma is proven.

**VI. IMPLEMENTATION DETAILS**

While the implementation of the first step is relatively simple, we need to specify some parameters and techniques that we used to implement raising 5 to an integer power.

The used computational precision \( p \) was equal to 128 bits. The standard C integer types give us either 32 or 64 bits, so for the implementation we used the \texttt{uint128_t} type from GCC that is realised with two 64-bit numbers. As a shifting parameter \( \lambda \) we took 64, so getting most or least 64 bits out of \texttt{uint128_t} number is easy and fast. Squarings and multiplications can be easily implemented using typecastings and appropriate shifts. Here, for instance, we put the code of squaring the 64-bit integer. The function returns two 64-bit integers, so the high and the low word of the 128-bit number.

```c
void square64(uint64_t *rh,
             uint64_t *rl,
             uint64_t a) {
    uint128_t r;
    r = ((uint128_t) a) * ((uint128_t) a);
    *rl = (uint64_t) r;
    r >>= 64;
    *rh = (uint64_t) r;
}
```

Listing 1. Example. C code sample for squaring a 64-bit number.
The other functions were implemented in the same manner.

We have implemented an run parametrized algorithm for computation of $5^B$, as the parameter we took the table index size (for entries $5^B$) and the working precision $p$. We see (Fig. 2) that the accuracy depends almost linearly on the precision.

VII. CONCLUSIONS

A novel algorithm for conversion between binary and decimal floating-point representations has been presented. All the computations are done in integer arithmetic, so no FP flags or modes can be influenced. This means that the corresponding code can be made reentrant. The exponent determination is exact and can be done with several basic arithmetic operations, stored constants and a table. The mantissa computation algorithm uses a small exact table. The error analysis is given and it corresponds to the experimental results. The accuracy of the result depends on the computing precision and the table size. The conversions are often used and the tables are multipurpose, so they can be reused by dozens of algorithms. As this conversion scheme is used everywhere and the tables are not large, they might be integrated in hardware. The implementation of the proposed algorithm can be done without loops, so it reduces the instructions that control the loop, optimizes and therefore accelerates the code. The described conversion approach was used in the implementation of the scanf analogue in libieee754 library [10].

REFERENCES