



# Parameterized floating-point logarithm and exponential functions for FPGAs

Jérémie Detrey, Florent de Dinechin

► **To cite this version:**

Jérémie Detrey, Florent de Dinechin. Parameterized floating-point logarithm and exponential functions for FPGAs. [Research Report] LIP RR-2006-02, Laboratoire de l'informatique du parallélisme. 2006, 2+13p. hal-02102238

**HAL Id: hal-02102238**

**<https://hal-lara.archives-ouvertes.fr/hal-02102238>**

Submitted on 17 Apr 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**Laboratoire de l'Informatique du Parallélisme**

École Normale Supérieure de Lyon  
Unité Mixte de Recherche CNRS-INRIA-ENS LYON-UCBL n° 5668

***Parameterized floating-point logarithm  
and exponential functions for FPGAs***

Jérémie Detrey ,  
Florent de Dinechin

January 2006

Research Report N° 2006-02

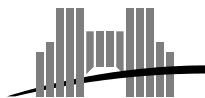
**École Normale Supérieure de Lyon**

46 Allée d'Italie, 69364 Lyon Cedex 07, France

Téléphone : +33(0)4.72.72.80.37

Télécopieur : +33(0)4.72.72.80.80

Adresse électronique : [lip@ens-lyon.fr](mailto:lip@ens-lyon.fr)



# Parameterized floating-point logarithm and exponential functions for FPGAs

J eremie Detrey , Florent de Dinechin

January 2006

## Abstract

As FPGAs are increasingly being used for floating-point computing, the feasibility of a library of floating-point elementary functions for FPGAs is discussed. An initial implementation of such a library contains parameterized operators for the logarithm and exponential functions. In single precision, those operators use a small fraction of the FPGA's resources, have a smaller latency than their software equivalent on a high-end processor, and provide about ten times the throughput in pipelined version. Previous work had shown that FPGAs could use massive parallelism to balance the poor performance of their basic floating-point operators compared to the equivalent in processors. As this work shows, when evaluating an elementary function, the flexibility of FPGAs provides much better performance than the processor without even resorting to parallelism. The presented library is freely available from <http://www.ens-lyon.fr/LIP/Arenaire/>.

**Keywords:** elementary functions, parameterized operators, logarithm, exponential, floating-point, FPGA, FPLibrary.

## R esum e

L'utilisation des FPGA pour des calculs en virgule flottante  tant de plus en plus courante, on peut d es lors envisager la r ealisation d'une biblioth eque de fonctions  l ementaires en virgule flottante pour FPGA. La premi ere version de cette biblioth eque propose des op erateurs param etr es pour le logarithme et l'exponentielle. En simple pr ecision, ces op erateurs utilisent une petite part des ressources du FPGA, pr esentent une latence inf erieure   celle de leur  quivalent logiciel sur un processeur actuel ainsi qu'un d ebit dix fois plus  lev e gr ace   l'utilisation du pipeline. Des travaux pr ec edents ont montr e que les FPGA pouvaient recourir au parall elisme pour pallier aux faibles performances des op erateurs virgule flottante classiques par rapport aux processeurs actuels. Comme le montre ce travail, en ce qui concerne l' valuation de fonctions  l ementaires, la souplesse des FPGA permet d'atteindre des performances bien meilleures qu'avec un processeur, et ce sans m eme tirer parti du parall elisme. La biblioth eque pr esent ee est librement accessible   l'adresse <http://www.ens-lyon.fr/LIP/Arenaire/>.

**Mots-cl es:** fonctions  l ementaires, op erateurs param etr es, logarithme, exponentielle, virgule flottante, FPGA, FPLibrary.

# 1 Introduction

A recent trend in FPGA computing is the increasing use of floating-point. Many libraries of floating-point operators for FPGAs now exist [18, 8, 1, 11, 6], typically offering the basic operators  $+$ ,  $-$ ,  $\times$ ,  $/$  and  $\sqrt{\phantom{x}}$ . Published applications include matrix operations, convolutions and filtering. As FPGA floating-point is typically clocked 10 times slower than the equivalent in contemporary processors, only massive parallelism (helped by the fact that the precision can match closely the application’s requirements) allows these applications to be competitive to software equivalent [13, 5, 10].

More complex floating-point computations on FPGAs will require good implementations of elementary functions such as logarithm, exponential, trigonometric, etc. These are the next useful building blocks after the basic operators. This paper describes both the logarithm and exponential functions, a first attempt to a library of floating-point elementary functions for FPGAs.

Elementary functions are available for virtually all computer systems. There is currently a large consensus that they should be implemented in software [17]. Even processors offering machine instructions for such functions (mainly the x86/x87 family) implement them as micro-code. On such systems, it is easy to design faster software implementations: Software can use large tables which wouldn’t be economical in hardware [19]. Therefore, no recent instruction set provides instructions for elementary functions.

Implementing floating-point elementary functions on FPGAs is a very different problem. The flexibility of the FPGA paradigm allows to use specific algorithms which turn out to be much more efficient than a processor-based implementation. We show in this paper that a single precision function consuming a small fraction of FPGA resources has a latency equivalent to that of the same function in a 2.4 GHz PC, while being fully pipelinable to run at 100 MHz. In other words, where the basic floating-point operator ( $+$ ,  $-$ ,  $\times$ ,  $/$ ,  $\sqrt{\phantom{x}}$ ) is typically 10 times slower on an FPGA than its PC equivalent, an elementary function will be more than ten times faster for precisions up to single precision.

Writing a *parameterized* elementary function is a completely new challenge: to exploit this flexibility, one should not use the same algorithms as used for implementing elementary functions in computer systems [19, 15, 14]. This paper describes an approach to this challenge, which builds upon previous work dedicated to fixed-point elementary function approximations (see [7] and references therein).

The authors are aware of only two previous works on floating-point elementary functions for FPGAs, studying the sine function [16] and studying the exponential function [9]. Both are very close to a software implementation. As they don’t exploit the flexibility of FPGAs, they are much less efficient than our approach, as section 4 will show.

## Notations

The input and output of our operators will be  $(3 + w_E + w_F)$ -bit floating-point numbers encoded in the freely available FPLibrary format [6] as follows:

- $F_X$ : The  $w_F$  least significant bits represent the fractional part of the mantissa  $M_X = 1.F_X$ .
- $E_X$ : The following  $w_E$ -bit word is the exponent, biased by  $E_0 = 2^{w_E-1} - 1$ .
- $S_X$ : The next bit is the sign of  $X$ .
- $\text{exn}_X$ : The two most significant bits of  $X$  are internal flags used to deal more easily with exceptional cases, as shown in Table 1.

exn <sub>X</sub>	X
00	0
01	$(-1)^{S_X} \cdot 1.F_X \cdot 2^{E_X - E_0}$
10	$(-1)^{S_X} \cdot \infty$
11	NaN ( <i>Not a Number</i> )

Table 1: Value of  $X$  according to its exception flags  $\text{exn}_X$ .

## 2 A floating-point logarithm

### 2.1 Evaluation algorithm

#### 2.1.1 Range reduction

We consider here only the case where  $X$  is a valid positive floating-point number (*ie.*  $\text{exn}_X = 01$  and  $S_X = 0$ ), otherwise the operator simply returns NaN. We therefore have:

$$X = 1.F_X \cdot 2^{E_X - E_0}.$$

If we take  $R = \log X$ , we obtain:

$$R = \log(1.F_X) + (E_X - E_0) \cdot \log 2.$$

In this case, we only have to compute  $\log(1.F_X)$  with  $1.F_X \in [1, 2)$ . The product  $(E_X - E_0) \cdot \log 2$  is then added back to obtain the final result.

In order to avoid catastrophic cancellation when adding the two terms, and consequently maintain low error bounds, we use the following equation to center the output range of the fixed-point log function around 0:

$$R = \begin{cases} \log(1.F_X) + (E_X - E_0) \cdot \log 2 & \text{when } 1.F_X \in [1, \sqrt{2}), \\ \log\left(\frac{1.F_X}{2}\right) + (1 + E_X - E_0) \cdot \log 2 & \text{when } 1.F_X \in [\sqrt{2}, 2). \end{cases} \quad (1)$$

We therefore have to compute  $\log M$  with the input operand  $M \in [\sqrt{2}/2, \sqrt{2})$ , which gives a result in the interval  $[-\log 2/2, \log 2/2)$ .

We also note in the following  $E = E_X - E_0$  when  $1.F_X \in [1, \sqrt{2})$ , or  $E = 1 + E_X - E_0$  when  $1.F_X \in [\sqrt{2}, 2)$ .

#### 2.1.2 Fixed-point logarithm

As we are targeting floating-point, we need to compute  $\log M$  with enough accuracy in order to guarantee faithful rounding, even after a possible normalization of the result. As  $\log M$  can be as close as possible to 0, a straightforward approach would require at least a precision of  $2w_F$  bits, as the normalization could imply a left shift of up to  $w_F$  bits, and  $w_F$  bits would still be needed for the final result.

But one can remark that when  $M$  is close to 1,  $\log M$  is close to  $M - 1$ . Therefore, a two-step approach consisting of first computing  $\log M / (M - 1)$  with a precision of  $w_F + g_0$  bits and then multiplying this result by  $M - 1$  (which is computed exactly) leads to the targeted accuracy at a smaller cost.

The function  $f(M) = \log M / (M - 1)$  is then computed by a generic polynomial method [7]. The order of the considered polynomial obviously depends on the precision  $w_F$ .

#### 2.1.3 Reconstruction

As the evaluation of  $f(M)$  is quite long, we can in parallel compute the sign of the result: If  $E = 0$ , then the sign will be the sign of  $\log M$ , which is in turn positive if  $M > 1$  and negative if  $M < 1$ . And if  $E \neq 0$ , as  $\log M \in [-\log 2/2, \log 2/2)$ , the sign will be the sign of  $E \cdot \log 2$ , which is the sign of  $E$ .

We can then compute in advance the opposite of  $E$  and  $M - 1$  and select them according to the sign of the result. Therefore, after the summation of the two products  $E \cdot \log 2$  and  $Y = f(M) \cdot (M - 1)$ , we obtain  $Z$  the absolute value of the result.

The last steps are of course the renormalization and rounding of this result, along with the handling of all the exceptional cases.

## 2.2 Architecture

The architecture of the logarithm operator is given on Figure 1. It is a straightforward implementation of the algorithm presented in Section 2.1. Due to its purely sequential dataflow, it can be easily pipelined. The values for the two parameters  $g_0$  and  $g_1$  are discussed in Section 2.3.

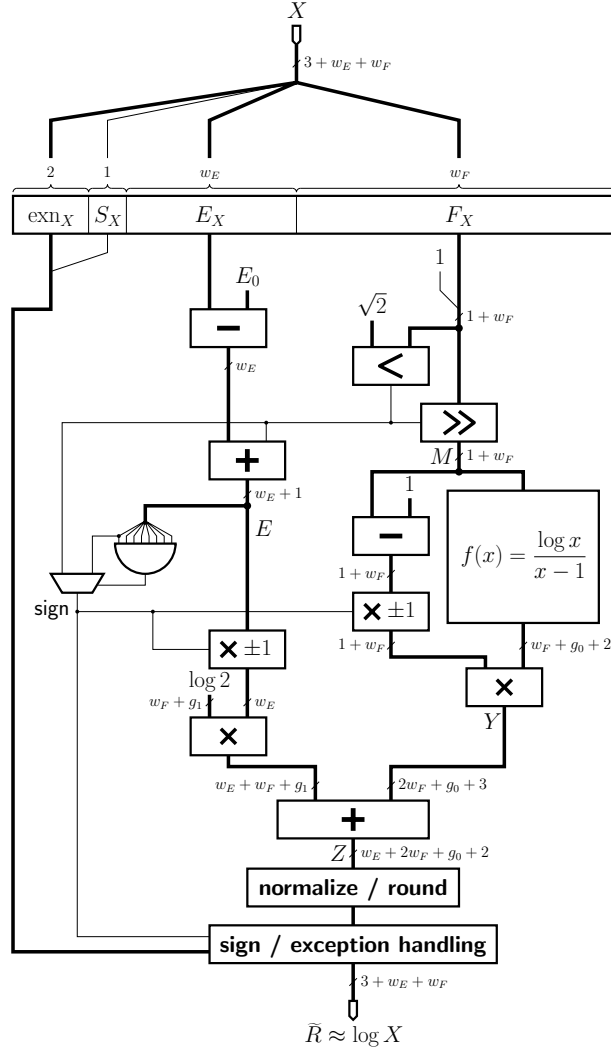


Figure 1: Architecture of the logarithm operator.

Some comments about this architecture:

- Remark that the boundary between the two cases of Equation (1) does not have to be exactly  $\sqrt{2}$ , as both alternatives are valid on the whole of  $[1, 2)$ . This means that the comparison between the mantissa  $1.F_X$  and  $\sqrt{2}$  may be performed on a few bits only, saving hardware. We do not have any longer that  $M \in [\sqrt{2}/2, \sqrt{2})$  and  $\log M \in [-\log 2/2, \log 2/2)$ , but we use the smallest approximation to  $\sqrt{2}$  that do not increase the bounds of these intervals to the

next power of two. Thus the savings in this step do not lead to increased hardware on the subsequent steps.

- The sign of the result is the sign of  $E$  when  $E \neq 0$ . If  $E = 0$ , we also need to take into account the result of the comparison between  $1.F_X$  and  $\sqrt{2}$ .
- The function  $f(x)$  is evaluated using the Higher-Order Table-Based Method (HOTBM) presented in [7]. It involves a piecewise polynomial approximation, with variable accuracy for the coefficients and where all the terms are computed in parallel.
- The normalization of the fixed-point number  $Z$  uses a leading-zero counter, and requires shifting  $Z$  by up to  $w_F$  bits on the left and up to  $w_E$  bits on the right.
- Underflow cases are detected by the *sign & exception handling* unit.

As the tables sizes grow exponentially with the precision, this architecture is well suited for precisions up to single precision ( $w_F = 23$  bits), and slightly more. Area on Virtex circuits will be given for a range of precision in Section 4.

### 2.3 Error analysis

In order to guarantee faithful rounding for the final result — *ie.* an error of less than one *unit in the last place* (ulp) of the result — we need to have a constant bound on the relative error of the fixed-point number  $Z = \log X$ :

$$\frac{|Z - \tilde{Z}|}{2^{\lfloor \log_2 |Z| \rfloor}} < 2^{-w_F - 1},$$

so that when rounding the result mantissa to the nearest, we obtain a total error bound of  $2^{-w_F}$ .

We need to consider several cases depending on the value of  $E$ :

- When  $|E| > 3$ ,  $|Z| > 2$  and the predominant error is caused by the discretization error of the  $\log 2$  constant, which is multiplied by  $E$ .
- When  $E = 0$ , on the other hand, the only error is caused by the evaluation of  $f(M)$ , which is then scaled in the product  $f(M) \cdot (M - 1)$ . As the multiplicand  $M - 1$  is computed exactly, this product does not entail any other error.
- When  $|E| = 2$  or  $3$ , both the discretization error from  $\log 2$  and the evaluation error from  $f(M)$  have to be taken into account. However, in this case, we have  $|Z| > 1$ . Therefore no cancellation will occur, and the discretization error will not be amplified by the normalization of  $Z$ .
- When  $|E| = 1$ , we have:

$$0.34 < \frac{1}{2} \log 2 \leq |Z| \leq \frac{3}{2} \log 2 < 1.04.$$

In this case, a cancellation of up to 2 bits can occur, which will multiply the  $\log 2$  discretization error by at most 4.

One can then find that using  $g_1 = 3$  guard bits for the  $\log 2$  constant and bounding the evaluation error  $\epsilon_f < 2^{-w_F - 3}$  satisfies all these constraints. The number of guard bits  $g_0$  is given by the evaluation scheme used for  $f(M)$ , and is typically comprised between 1 and 5 bits.

All these choices have been proven valid by exhaustively testing our operators on a Celoxica RC-1000 board (with a VirtexE-2000 FPGA) against a double precision software function, for the whole parameter space defined by  $w_E \in [3, 8]$  and  $w_F \in [6, 23]$ . This exhaustive testing showed that the the result was always faithful, and was correctly rounded to nearest in more than 98% of the cases.

## 3 A floating-point exponential

### 3.1 Evaluation algorithm

#### 3.1.1 Special cases

The exponential function is defined on the set of the reals. However, in this floating-point format, the smallest representable number is:

$$X_{\min} = 2^{1-E_0},$$

and the largest is:

$$X_{\max} = \left(1 + \frac{2^{w_F} - 1}{2^{w_F}}\right) \cdot 2^{2^{w_E} - 2 - E_0}.$$

The exponential should return zero for all input numbers smaller than  $\log(X_{\min})$ , and should return  $+\infty$  for all input numbers larger than  $\log(X_{\max})$ . In single precision ( $w_E = 8$ ,  $w_F = 23$ ), for instance, the set of input numbers on which a computation will take place is  $[-87.34, 88.73]$ . The remainder of this section only describes the computation on this interval.

#### 3.1.2 A first algorithm

The straightforward idea to compute the exponential of  $X$  is to use the identity:

$$e^X = 2^{X/\log(2)}.$$

Therefore, first compute  $X/\log(2)$  as a fixed-point value  $Y = Y_{\text{int}}.Y_{\text{frac}}$ . The integer part of  $Y$  is then the exponent of the exponential of  $X$ , and to get the fraction of  $e^X$  one needs to compute the function  $2^x$  with the fractional part of  $Y$  as input.

This approach poses several problems:

- To be sure of the exponent, one has to compute  $Y$  with very good accuracy: A quick error analysis shows that one needs to use a value of  $\frac{1}{\log(2)}$  on more than  $w_E + w_F$  bits, which in practice means a very large multiplier for the computation of  $X \cdot \frac{1}{\log(2)}$ .
- The accurate computation of  $2^{Y_{\text{frac}}}$  will be very expensive as well. Using a table-based method, it needs a table with at least  $w_F$  bits of inputs.

The second problem can be solved using a second range reduction, splitting  $Y_{\text{frac}}$  into two subwords:

$$Y_{\text{frac}} = Y_1 + Y_2,$$

where  $Y_1$  holds the  $p$  most significant bits of  $Y$ . One may then use the identity:

$$2^{Y_1+Y_2} = 2^{Y_1} \cdot 2^{Y_2}.$$

But again, we would have a multiplier of size at least  $w_F \times (w_F - p)$ .

#### 3.1.3 Improved algorithm

A slightly more complicated algorithm, closer to what is typically used in software [12], solves the previous problems. The main idea is to reduce  $X$  to an integer  $k$  and a fixed-point number  $Y$  such as:

$$X \approx k \cdot \log(2) + Y, \tag{2}$$

and then to use the identity:

$$e^X \approx 2^k \cdot e^Y.$$



The reduction to  $(k, Y)$  is implemented by first computing  $k \approx X \cdot \frac{1}{\log(2)}$ , then computing  $Y = X - k \cdot \log(2)$ . The computation of  $e^Y$  may use a second range reduction as previously, splitting  $Y$  as:

$$Y = Y_1 + Y_2,$$

where  $Y_1$  holds the  $p$  most significant bits of  $Y$ , then computing:

$$e^{Y_1+Y_2} = e^{Y_1} \cdot e^{Y_2},$$

where  $e^{Y_1}$  will be tabulated.

This looks similar to using the naive approach, however it has two main advantages: First, the computation of  $k$  can be approximated, as long as the computation of  $Y$  compensates it in such a way that Equation (2) is accurately implemented. As  $k$  is a small integer, this in practice replaces a large multiplication with two much smaller multiplications, one to compute  $k$ , the second to compute  $k \cdot \log(2)$ . This approximation, however, implies that the final exponent may be  $k \pm 1$ : the result  $2^k \cdot e^Y$  will require a normalization.

Second, computing  $e^{Y_2}$  is simpler than computing  $2^{Y_2}$ , because of the Taylor formula:

$$e^{Y_2} \approx 1 + Y_2 + T(Y_2),$$

where  $T(Y_2) \approx e^{Y_2} - 1 - Y_2$  will be also tabulated.

This not only reduces a table-based approximation to a much smaller one (as  $Y_2 < 2^{-p-1}$  as will be seen in Section 3.3, it requires about  $w_F - 2p$  input bits instead of  $w_F - p$  bits), it also offers the opportunity to save  $p$  lines of the final large multiplier by implementing it as:

$$e^{Y_1} \cdot (1 + Y_2 + T(Y_2)) = e^{Y_1} + e^{Y_1} \cdot (Y_2 + T(Y_2)).$$

The relevance of this depends on the target architecture. Obviously, it is relevant to FPGAs without embedded multipliers. If such multipliers are available (like the  $18 \times 18$  of some Virtex architectures), it is relevant if the size of one of the inputs gets smaller than 18. For instance a  $18 \times 24$  multiplication followed by one addition may be considered more economical than a  $24 \times 24$  multiplication consuming 4 embedded multipliers (see a detailed study of multiplier implementation on the Virtex family in [2]). Conversely, if the rectangular multiplication consumes 4 embedded multipliers anyway, the addition should also be computed by these multipliers. This is the case for single precision.

### 3.1.4 A table-driven method

The previous algorithm involves two tables:

- The first, with  $p$  input bits and a few more than  $w_F$  output bits for  $e^{Y_1}$ , can not be compressed using table-based methods derived from the bipartite method.
- The second, with about  $w_F - 2p$  input bits and as many output bits, can. As for the logarithm operator, we use the HOTBM method [7].

Compressed or not, the sizes of these tables grow exponentially with their input size. Just like the logarithm, this algorithm is therefore well suited for precisions up to (and slightly above) single precision ( $w_F = 23$  bits). Area on Virtex circuits will be given for a range of precision in Section 4.

For much smaller precisions, of course, simpler approaches will be more effective. For much larger precisions, like double precision, the algorithm has to be modified. An idea is to repeat the second range reduction several times, each time replacing  $p$  input bits to the tables by one new  $p$ -input-bits table and one almost full-size multiplier. Another solution is to compute  $e^{Y_2}$  using a higher degree polynomial, which also increases the multiplier count. A more detailed study remains to be done.

### 3.2 Architecture

The architecture of this operator is given on Figure 2. It is a straightforward implementation of the algorithm. Obviously this architecture is purely sequential and can be pipelined easily.

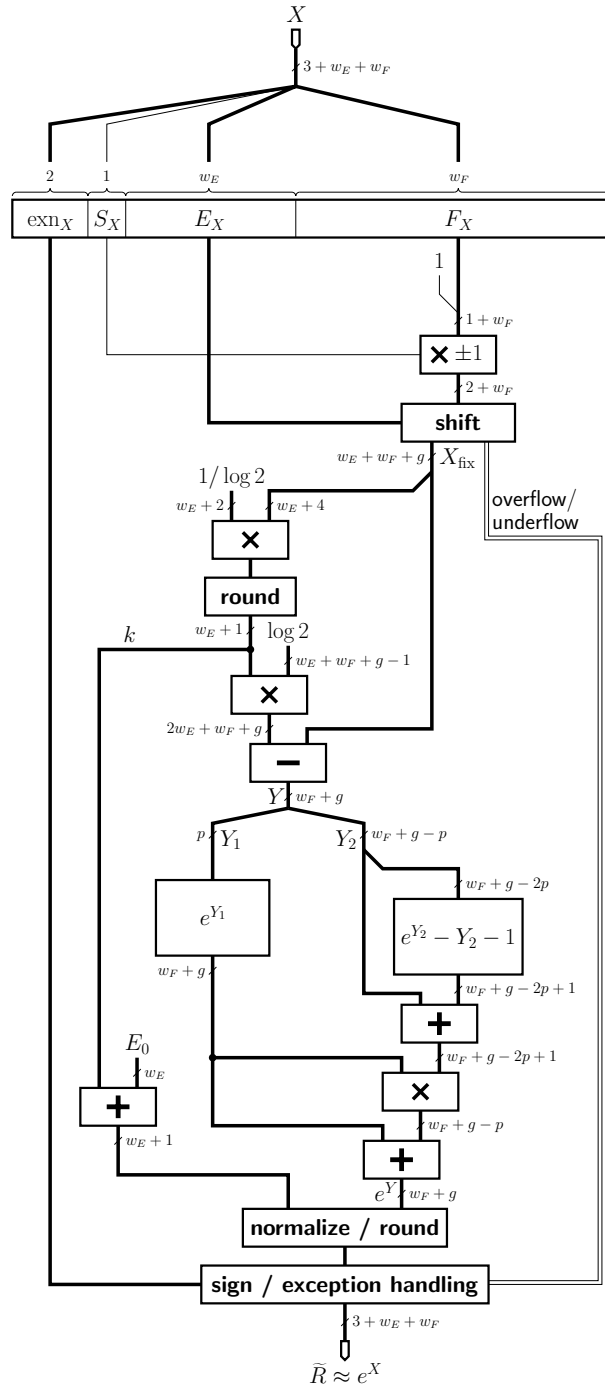


Figure 2: Architecture of the exponential operator.

The architecture has two parameters,  $p$  and  $g$ . The first essentially drives a tradeoff between the sizes of the two tables, and its value should be comprised between  $p = w_F/4$  and  $p = w_F/3$ . The second is a number of guard bits used to contain rounding errors, and will be studied in

### Section 3.3.

Some more comments about this architecture:

- The *shift* operator shifts the mantissa by the value of the exponent. More specifically, if the exponent is positive, it shifts to the left by up to  $w_E$  positions (more would mean overflow). If the exponent is negative, it shifts to the right by up to  $w_F + g$  positions. The result is then truncated to  $w_E + w_F + g$  bits.
- The range check (which verifies if the exponential of the input is representable in the given format, or if an infinity or a zero should be returned) is performed by the *shift* and the first multiplication stages.
- The intermediate value of  $X_{\text{fix}}$  has  $w_E + w_F + g + 1$  bits with a fixed binary point after the  $w_E + 1$ -th. The computation of  $X_{\text{fix}} - k \cdot \log(2)$  will cancel the integer part and the first bit of the fractional part, as shown below in Section 3.3.
- The first two multipliers are constant multipliers, for which a range of optimization techniques may apply [3, 4]. This is currently not exploited.
- This figure shows the final multiplication implemented as a multiplier followed by an adder, but as shown in Section 3.1.3, depending on the target architecture, it may make more sense to have one single multiplier instead.
- Final normalization possibly shifts left the mantissa by one bit (as will be shown in 3.3.1), then rounds the mantissa, then possibly shifts back right by one bit in the rare case when rounding also changes the exponent. Each shift is associated with an increment/decrement of the exponent.

Several of these blocks can certainly be the subject of further minor optimizations.

## 3.3 Error analysis

The goal is to obtain a floating-point operator which guarantees faithful rounding. There are two aspects to the error analysis. First, the range reduction should be implemented with the minimum hardware. Second, the whole of the computation should ensure faithful rounding, considering the method and rounding errors involved.

### 3.3.1 Range reduction

As  $k$  will be the exponent (plus or minus 1) of the final exponential, it fits on a  $w_E + 1$  machine word.

If the range reduction step were exact, the value of  $Y$  would be in  $[-\frac{\log(2)}{2}, \frac{\log(2)}{2}]$ , ensuring that  $e^Y$  is in  $[\frac{\sqrt{2}}{2}, \sqrt{2}]$ . Using less accuracy to compute  $k$ , we accept that  $Y$  will be in an interval larger than  $[-\frac{\log(2)}{2}, \frac{\log(2)}{2}]$ . It will make little difference to the architecture to increase these bounds to the next power of two, which is  $Y \in ]-1/2, 1/2[$ . One proves easily that this is obtained for all  $w_E$  by truncating  $1/\log(2)$  to  $w_E + 1$  bits, and considering only  $w_E + 3$  bits of  $X_{\text{fix}}$ .

This means that we will have  $e^Y$  in  $[0.6, 1.7]$ , so we will sometimes have to normalize the final result by shifting the mantissa one bit to the right and increasing the exponent by one.

### 3.3.2 Faithful rounding

The computation of  $e^Y$  involves a range of approximation and rounding errors, and the purpose of this section is to guarantee faithful rounding with a good percentage of correct rounding.

In the following, the errors will be expressed in terms of units in the last place (ulps) of  $Y$ . It is safe to reason in terms of ulps since all the computations are in fixed point, which makes it easy to align the binary point of each intermediate value. Here the ulp has the value  $2^{-w_F - g}$ . Then we can make an error expressed this way as small as required by increasing  $g$ .

First, note that the argument reduction is not exact. It entails an error due to:

- the approximation of  $\log(2)$  to  $w_E + w_F + g - 1$  bits (less than one half-ulp),
- $X_{\text{fix}}$  which is exact if it was shifted left, but was truncated if shifted right (one ulp),
- the truncation of  $Y$  to  $w_F + g$  bits (one ulp).

Thus in the worst case we have lost 5 half-ulps.

Now we consider subsequent computations on  $Y$  carrying this error.

The table of  $e^{Y_1}$  holds an error of at most one half-ulp.

The table of  $e^{Y_2} - Y_2 - 1$  is only faithful because it uses the HOTBM compression technique (error up to one ulp, plus another ulp when truncating  $Y_2$  to its most significant part). The previous error on  $Y$  is negligible for this table as its result is scaled by  $2^{-2p-1}$ .

Due to the multiplier, the error due to the second table (2 ulps) added to the error on  $Y_2$  (5 half-ulps) may be scaled by the value contained in the first table (less than 1.7). This leads to an error of less than 8 ulps.

The first addition involves no error, we again lose one half-ulp when rounding the result of the multiplication, and the second addition adds the half-ulp error from the first table.

Finally the errors sum up to 9 ulps. Besides we have to take into account that we may need to shift the mantissa left in case of renormalization, so we have to provide one extra bit of accuracy for that. Altogether, we find that  $g = 5$  guard bits for the intermediate computations ensure faithful rounding.

A finer error analysis directing slight modifications of the algorithm (replacing some of the truncations by roundings) could probably reduce  $g$ , but would also increase the critical path.

As for the logarithm operator, we implemented a test procedure which compares the result of this operator on a Celoxica RC-1000 board against a double precision exponential on the host PC. Exhaustive testing for various precision has confirmed that the result is always faithful, and correctly rounded to nearest in more than 75% of the cases.

## 4 Results

We obtained area and delay estimations of our operators for several precisions. These results were computed using Xilinx ISE and XST 6.3 for a Virtex-II XC2V1000-4 FPGA. They are shown in Figure 3, and a summary is given in Table 2, in terms of slices and percentage of FPGA occupation for the area, and in terms of nanoseconds for the latency.

Precision ( $w_E, w_F$ )	Multipliers	Logarithm			Exponential			
		Area (slices % mults)	Latency (ns)		Area (slices % mults)	Latency (ns)		
(3,6)	LUT-based	123 (2%)	–	34	137 (2%)	–	51	
	$18 \times 18$	89 (1%)	2	31	68 (1%)	3	47	
(5,10)	LUT-based	263 (5%)	–	42	258 (5%)	–	63	
	$18 \times 18$	154 (3%)	3	39	135 (2%)	4	57	
(6,13)	LUT-based	411 (8%)	–	48	357 (6%)	–	69	
	$18 \times 18$	233 (4%)	3	44	194 (3%)	5	65	
(7,16)	LUT-based	619 (12%)	–	57	480 (9%)	–	69	
	$18 \times 18$	343 (6%)	6	55	271 (5%)	5	68	
(8,23)	LUT-based	1399 (27%)	–	64	948 (18%)	–	85	
	$18 \times 18$	830 (16%)	9	61	522 (10%)	9	83	

Table 2: Synthesis results for the operators on Xilinx Virtex-II.

In order to be as portable as possible, we do not require the use of the specific Virtex-II embedded  $18 \times 18$  multipliers. Therefore we present the results obtained with and without those multipliers in Figure 4.

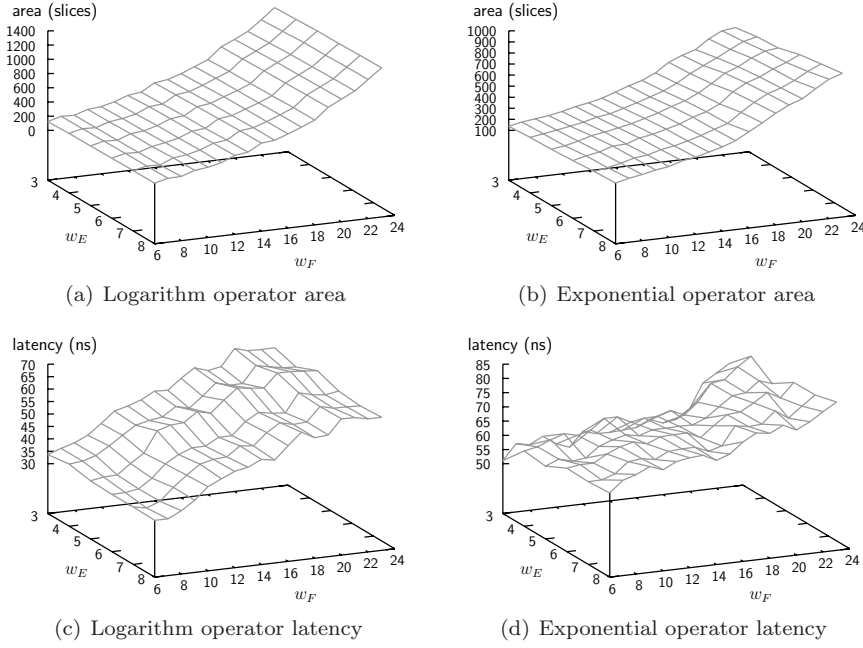


Figure 3: Area and latency estimations depending on  $w_E$  and  $w_F$  for the combinatorial operators with LUT-based multipliers.

If most of the results presented here are for the combinatorial version, our operators are also available as pipelined operators, for a small overhead in area, as shown in Figure 5. The pipeline depth depends on the parameters  $w_E$  and  $w_F$ : between 5 and 11 cycles for the logarithm, and from 10 to 15 cycles for the exponential operator. The pipelined operators are designed to run at 100 MHz on the targeted Virtex-II XC2V1000-4 FPGA.

As a comparison, Table 3 presents the performances for both our operators in single precision, along with the measured performances for a 2.4 GHz Intel Xeon processor, using the single precision operators from the GNU `glibc` (which themselves rely on the micro-coded machine instructions `fyl2x`, `fyl2xp1` and `f2xm1`).

Function	2.4 GHz Intel Xeon			100 MHz Virtex-II FPGA		
	Cycles	Latency (ns)	Throughput ( $10^6$ op/s)	Cycles	Latency (ns)	Throughput ( $10^6$ op/s)
Logarithm	196	82	12	11	64	100
Exponential	308	128	8	15	85	100

Table 3: Performance comparison between Intel Xeon and Virtex-II for single precision.

The only other comparable work we could find in the literature [9] reports 5564 slices for a single precision exponential unit which computes exponentials in 74 cycles fully pipelined at 85 MHz on a Virtex-II 4000. Our approach is much more efficient, because our algorithm is designed from scratch specifically for the FPGA. In contrast, the authors of [9] use an algorithm designed for microprocessors. In particular, they internally use fully featured floating-point adders and multipliers everywhere where we only use fixed-point operators.

## 5 Conclusion and future work

Parameterized floating-point implementations for the logarithm and exponential functions have been presented. For the 32-bit single precision format, their latency matches that of a Xeon

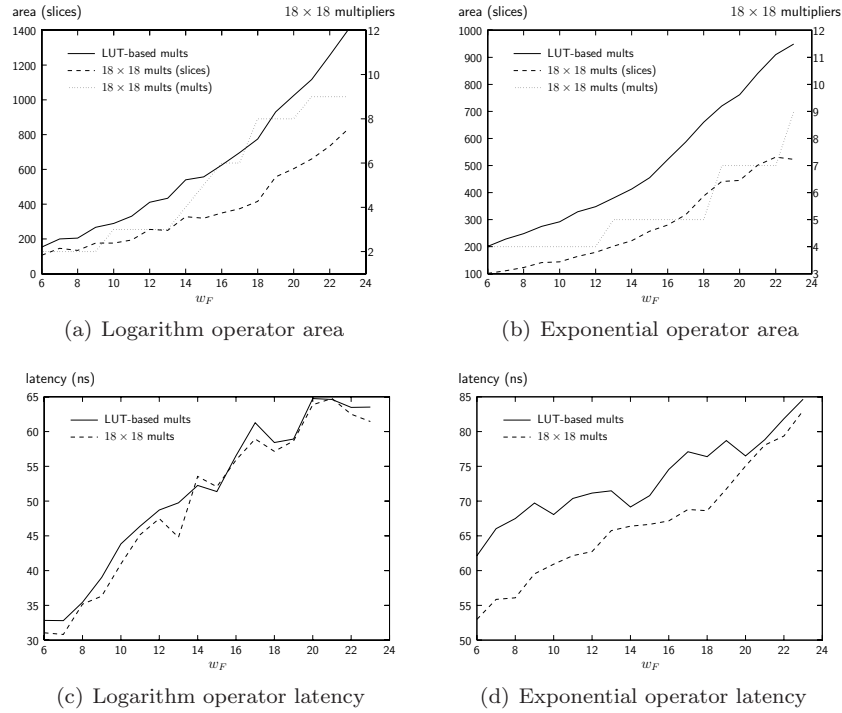


Figure 4: Comparison of area and latency depending on  $w_F$  ( $w_E = 8$ ), when using LUT-based multipliers, and when using the embedded  $18 \times 18$  multipliers.

processor, and their pipelined version provide several times the Xeon throughput. Besides, they consume a small fraction of the FPGA’s resources.

We should moderate these results by a few remarks. Firstly, our implementations are slightly less accurate than the Xeon ones, offering faithful rounding only, where the Xeon uses an internal precision of 80 bits which ensures almost guaranteed correct rounding. Secondly, more recent instruction sets allow for lower latency for the elementary functions. The Itanium 2, for example, can evaluate a single precision exponential in about 40 cycles (or 20 ns at 2 GHz), and will therefore be just twice slower than our pipelined implementation. Thirdly, implementations for the logarithm or the exponential better optimized for single precision could probably be written for these recent processors. However the argument of massive parallelism will still apply.

A future research direction, already evoked, is that the current architectures do not scale well beyond single precision: some of the building blocks have a size exponential in the precision. We will therefore explore algorithms which work up to double precision, which is the standard in processors - and soon in FPGAs [5, 10]. We are also investigating other elementary functions to extend the library.

Another future goal is to study the portability and applicability of these algorithms to ASICs (Application-Specific Integrated Circuits), where the metrics are very different and offer new trade-offs which should be investigated. For example, as look-up tables implemented as ROMs can be very area-efficient in ASIC technology, one could consider increasing the size of the tables in the HOTBM operator to reduce the overall complexity of the complete operators.

FPLibrary and the operators presented here are available under the GNU Public Licence from <http://www.ens-lyon.fr/LIP/Arenaire/>.

## Acknowledgements

The authors would like to thank Arnaud Tisserand for many interesting discussions and for maintaining the servers and software on which the experiments were conducted.

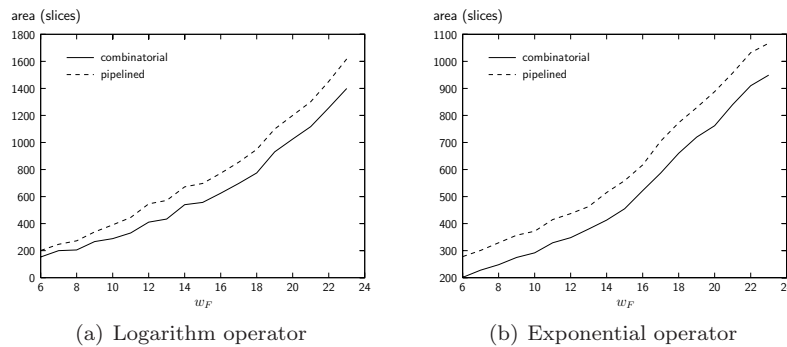


Figure 5: Area estimations depending on  $w_F$  ( $w_E = 8$ ) for the combinatorial and pipelined versions of the operators with LUT-based multipliers.

## References

- [1] P. Belanović and M. Leaser. A library of parameterized floating-point modules and their use. In *Field Programmable Logic and Applications*, volume 2438 of *LNCS*, pages 657–666. Springer, Sept. 2002.
- [2] J.-L. Beuchat and A. Tisserand. Small multiplier-based multiplication and division operators for Virtex-II devices. In *Field-Programmable Logic and Applications*, volume 2438 of *LNCS*, pages 513–522. Springer, Sept. 2002.
- [3] K. Chapman. Fast integer multipliers fit in FPGAs (EDN 1993 design idea winner). *EDN magazine*, May 1994.
- [4] F. de Dinechin and V. Lefèvre. Constant multipliers for FPGAs. In *2nd Intl Workshop on Engineering of Reconfigurable Hardware/Software Objects (ENREGLE)*, pages 167–173, June 2000.
- [5] M. deLorimier and A. DeHon. Floating-point sparse matrix-vector multiply for FPGAs. In *ACM/SIGDA Field-Programmable Gate Arrays*, pages 75–85. ACM Press, 2005.
- [6] J. Detrey and F. de Dinechin. A tool for unbiased comparison between logarithmic and floating-point arithmetic. Technical Report RR2004-31, LIP, École Normale Supérieure de Lyon, Mar. 2004.
- [7] J. Detrey and F. de Dinechin. Table-based polynomials for fast hardware function evaluation. In *16th Intl Conference on Application-specific Systems, Architectures and Processors*. IEEE Computer Society Press, July 2005.
- [8] J. Dido, N. Geraudie, L. Loiseau, O. Payeur, Y. Savaria, and D. Poirier. A flexible floating-point format for optimizing data-paths and operators in FPGA based DSPs. In *ACM/SIGDA Field-Programmable Gate Arrays*, pages 50–55, Feb. 2002.
- [9] C. Doss and R. Riley. FPGA-based implementation of a robust IEEE-754 exponential unit. In *FPGAs for Custom Computing Machines*. IEEE, 2004.
- [10] Y. Dou, S. Vassiliadis, G. K. Kuzmanov, and G. N. Gaydadjiev. 64-bit floating-point FPGA matrix multiplication. In *ACM/SIGDA Field-Programmable Gate Arrays*, pages 86–95. ACM Press, 2005.
- [11] B. Lee and N. Burgess. Parameterisable floating-point operators on FPGAs. In *36th Asilomar Conference on Signals, Systems, and Computers*, pages 1064–1068, 2002.

- [12] R.-C. Li, P. Markstein, J. P. Okada, and J. W. Thomas. The libm library and floating-point arithmetic for HP-UX on Itanium. Technical report, Hewlett-Packard company, april 2001.
- [13] G. Lienhart, A. Kugel, and R. Männer. Using floating-point arithmetic on FPGAs to accelerate scientific N-body simulations. In *FPGAs for Custom Computing Machines*. IEEE, 2002.
- [14] P. Markstein. *IA-64 and Elementary Functions: Speed and Precision*. Hewlett-Packard Professional Books. Prentice Hall, 2000. ISBN: 0130183482.
- [15] J.-M. Muller. *Elementary Functions, Algorithms and Implementation*. Birkhauser, Boston, 1997.
- [16] F. Ortiz, J. Humphrey, J. Durbano, and D. Prather. A study on the design of floating-point functions in FPGAs. In *Field Programmable Logic and Applications*, volume 2778 of *LNCIS*, pages 1131–1135. Springer, Sept. 2003.
- [17] G. Paul and M. W. Wilson. Should the elementary functions be incorporated into computer instruction sets? *ACM Transactions on Mathematical Software*, 2(2):132–142, June 1976.
- [18] N. Shirazi, A. Walters, and P. Athanas. Quantitative analysis of floating point arithmetic on FPGA based custom computing machine. In *FPGAs for Custom Computing Machines*, pages 155–162. IEEE, 1995.
- [19] P. T. P. Tang. Table lookup algorithms for elementary functions and their error analysis. In P. Kornerup and D. W. Matula, editors, *Proceedings of the 10th IEEE Symposium on Computer Arithmetic*, pages 232–236, Grenoble, France, June 1991. IEEE.