Introduction
Floating point numbers
Steps to calculate crlog
Algorithm
Conclusion

Logarithm with correct rounding

Sid-Ali Zitouni-Terki

September 8th 2022



Outline

- Introduction
- Ploating point numbers
 - Definition
 - Necessary tools
- Steps to calculate crlog
- 4 Algorithm
 - Operator precision calculation
 - Logarithme
- Conclusion



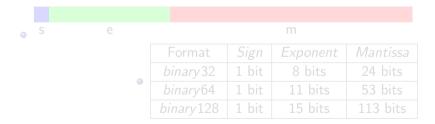
Introduction Floating point numbers Steps to calculate crlog Algorithm Conclusion

Introduction

Logarithm with correct rounding

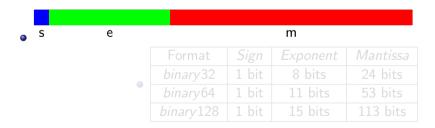
Floating point numbers

• $x = s.M.\beta^{e-p+1}$ with s the sign, M the mantissa ($|M| \le \beta^p - 1$), β a base(radix), e the exponent and p the precision.



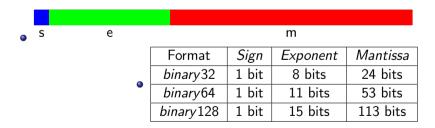
Floating point numbers

• $x = s.M.\beta^{e-p+1}$ with s the sign, M the mantissa ($|M| \le \beta^p - 1$), β a base(radix), e the exponent and p the precision.



Floating point numbers

• $x = s.M.\beta^{e-p+1}$ with s the sign, M the mantissa ($|M| \le \beta^p - 1$), β a base(radix), e the exponent and p the precision.



- RNDN:
 - Ties to even
 - Ties to away
- RNDZ
- RNDU
- RNDD

Real Number	+140.8215064611465	+665.5752955525412
RNDN(Ties to even)	+140.821506461146	+665.575295552541
RNDN(Ties to away)	+140.821506461147	+665.575295552541
RNDZ	+140.821506461146	+665.575295552541
RNDU	+140.821506461147	+665.575295552542
RNDD	+140.821506461146	+665.575295552541

- RNDN:
 - Ties to even
 - Ties to away
- RNDZ
- RNDU
- RNDD

Real Number	+140.8215064611465	+665.5752955525412
RNDN(Ties to even)	+140.821506461146	+665.575295552541
RNDN(Ties to away)	+140.821506461147	+665.575295552541
RNDZ	+140.821506461146	+665.575295552541
RNDU	+140.821506461147	+665.575295552542
RNDD	+140.821506461146	+665.575295552541

- RNDN:
 - Ties to even
 - Ties to away
- RNDZ
- RNDU
- RNDD

Real Number	+140.8215064611465	+665.5752955525412
RNDN(Ties to even)	+140.821506461146	+665.575295552541
RNDN(Ties to away)	+140.821506461147	+665.575295552541
RNDZ	+140.821506461146	+665.575295552541
RNDU	+140.821506461147	+665.575295552542
RNDD	+140.821506461146	+665.575295552541

- RNDN:
 - Ties to even
 - Ties to away
- RNDZ
- RNDU
- RNDD

Real Number	+140.8215064611465	+665.5752955525412
RNDN(Ties to even)	+140.821506461146	+665.575295552541
RNDN(Ties to away)	+140.821506461147	+665.575295552541
RNDZ	+140.821506461146	+665.575295552541
RNDU	+140.821506461147	+665.575295552542
RNDD	+140.821506461146	+665.575295552541

- RNDN:
 - Ties to even
 - Ties to away
- RNDZ
- RNDU
- RNDD

•

Real Number	+140.8215064611465	+665.5752955525412
RNDN(Ties to even)	+140.821506461146	+665.575295552541
RNDN(Ties to away)	+140.821506461147	+665.575295552541
RNDZ	+140.821506461146	+665.575295552541
RNDU	+140.821506461147	+665.575295552542
RNDD	+140.821506461146	+665.575295552541



- \bullet +0 and -0
- \bullet $-\infty$ and $+\infty$
- NaN
- Subnormal
- Overflow
- Underflow

Definition (ulp(x))

Let x a **floating-point** number, $x=d_0d_1d_2d_3d_4...d_{p-1}\beta^e$. we therefore have an error to represent it : $|d_0d_1d_2d_3d_4...d_{p-1}-\frac{x}{\beta^e}|$ which is the unit of the last place.

•

Corollary

If x is a **floating-point** number, $d_0 = 1$ and that its base is 2, then we have

$$ulp(x) \le 2^{-52}|x|$$

Definition (ulp(x))

Let x a **floating-point** number, $x = d_0d_1d_2d_3d_4...d_{p-1}\beta^e$. we therefore have an error to represent it : $|d_0d_1d_2d_3d_4...d_{p-1} - \frac{x}{\beta^e}|$ which is the unit of the last place.

•

Corollary

If x is a **floating-point** number, $d_0 = 1$ and that its base is 2, then we have :

$$ulp(x) \le 2^{-52}|x|$$

Property

Property

If
$$X = RNDN(x) \Rightarrow |X - x| \le \frac{1}{2}ulp(x)$$

Property

If
$$X = RNDN(x) \Rightarrow |X - x| \leq \frac{1}{2}ulp(X)$$

Property

If
$$X \in \{RNDD(x), RNDU(x), RNDZ(x), RNDN(x)\} \Rightarrow |X - x| \leq ulp(x)$$

Property

If $X \in \{RNDD(x), RNDU(x), RNDZ(x), RNDN(x)\} \Rightarrow |X - x| \leq ulp(X)$

Unit Roundoff

$$u = \begin{cases} \frac{1}{2}.ulp(1) = \frac{1}{2}.2^{-52} = 2^{-53} & \text{for } RNDN \\ ulp(1) = 2^{-52} & \text{for } (RNDU, RNDZ, RNDD) \end{cases}$$

Sterbenz's lemma

Lemma (Sterbenz)

In a **radix**- β floating-point system with **subnormal** numbers available, if x and y are finite **floating-point** numbers such that $\frac{y}{2} \le x \le 2.y$, then x-y is exactly representable.

Correct rounding methodology

- $cr \log_{fast}$ computes an approximation y of f(x) with a small error bound ϵ_1 ;
- ② rounding test : $[y \epsilon_1, y + \epsilon_1]$
- **3** $cr \log accurate$ computes an approximation y of f(x) with a small error bound ϵ_2 with $|\epsilon_2| < |\epsilon_1|$
- **o** rounding test : $[y \epsilon_2, y + \epsilon_2]$
- or log advanced give us the correctly rounded value.

Fast path

- Special cases
- Argument reduction
- Opening approximation and evaluation
- Result construction

Special cases

- input is NaN
- Input is negatif
- Input is +0 or -0
- Input is $+\infty$
- Input is a subnormal number. We transform it into normal number then we calculate as if it is normal.

- $x = 2^E \times m$
- Condition before to use Tang's algorithm: 1 < m < 21.b.....b with $b \in \{0,1\}$
- k significant bits after the initial 1, i = Integer(k.bits)
 - 1. b..b b with $b \in \{0, 1\}$
 - k bits
- $\alpha_i = \frac{2^k}{2^k + i}$
- $m = \frac{m \times \alpha_i}{\alpha_i} \Rightarrow \log(m) = \log(m * \alpha_i) \log(\alpha_i)$
- $1 < m' < 1 + \frac{1}{2^k}$ with $m' = m \times \alpha_i$
 - 1. 0..0b with $b \in \{0, 1\}$

- $x = 2^E \times m$
- $log(x) = E \times log(2) + log(m)$
- Condition before to use Tang's algorithm: 1 < m < 2 1.b......b with $b \in \{0,1\}$
- k significant bits after the initial 1, i = Integer(k.bits)1. b..bb with $b \in \{0, 1\}$
 - k bits
- $\alpha_i = \frac{2^k}{2^k + i}$
- $m = \frac{m \times \alpha_i}{\alpha_i} \Rightarrow \log(m) = \log(m * \alpha_i) \log(\alpha_i)$
- $1 < m' < 1 + \frac{1}{2^k}$ with $m' = m \times \alpha_i$
 - 1. 0..0b with $b \in \{0, 1\}$

- $x = 2^E \times m$
- Condition before to use Tang's algorithm: 1 < m < 2 1.b......b with $b \in \{0,1\}$
- k significant bits after the initial 1, i = Integer(k.bits)1. $\underbrace{b..b}_{k\ bits}$ b with $b \in \{0,1\}$
- $\bullet \ \alpha_i = \frac{2^k}{2^k + i}$
- $m = \frac{m \times \alpha_i}{\alpha_i} \Rightarrow \log(m) = \log(m * \alpha_i) \log(\alpha_i)$
- $1 < m' < 1 + \frac{1}{2^k}$ with $m' = m \times \alpha_i$ 1. 0..0b with $b \in \{0, 1\}$

- $x = 2^E \times m$
- Condition before to use Tang's algorithm: 1 < m < 2 1.b......b with $b \in \{0,1\}$
- k significant bits after the initial 1, i = Integer(k.bits)1. $\underbrace{b..b}_{k\ bits}$ b with $b \in \{0,1\}$

$$\alpha_i = \frac{2^k}{2^k + i}$$

•
$$m = \frac{m \times \alpha_i}{\alpha_i} \Rightarrow \log(m) = \log(m * \alpha_i) - \log(\alpha_i)$$

•
$$1 < m' < 1 + \frac{1}{2^k}$$
 with $m' = m \times \alpha_i$
1. 0..0b with $b \in \{0, 1\}$

- $x = 2^E \times m$
- $log(x) = E \times log(2) + log(m)$
- Condition before to use Tang's algorithm: 1 < m < 2 1.b......b with $b \in \{0,1\}$
- k significant bits after the initial 1, i = Integer(k.bits)1. $\underbrace{b..b}_{k\ bits}$ b with $b \in \{0,1\}$
- $\bullet \ \alpha_i = \frac{2^k}{2^k + i}$
- $m = \frac{m \times \alpha_i}{\alpha_i} \Rightarrow \log(m) = \log(m * \alpha_i) \log(\alpha_i)$
- $1 < m' < 1 + \frac{1}{2^k}$ with $m' = m \times \alpha_i$ 1. 0..0b with $b \in \{0, 1\}$

- $x = 2^E \times m$
- Condition before to use Tang's algorithm: 1 < m < 2 1.b......b with $b \in \{0,1\}$
- k significant bits after the initial 1, i = Integer(k.bits)1. b..bb with $b \in \{0,1\}$
 - k bits
- $\bullet \ \alpha_i = \frac{2^k}{2^k + i}$
- $m = \frac{m \times \alpha_i}{\alpha_i} \Rightarrow \log(m) = \log(m * \alpha_i) \log(\alpha_i)$
- $1 < m' < 1 + \frac{1}{2^k}$ with $m' = m \times \alpha_i$ 1. 0..0b with $b \in \{0, 1\}$

- $x = 2^E \times m$
- $log(x) = E \times log(2) + log(m)$
- Condition before to use **Tang**'s algorithm: 1 < m < 2 1.b.....b with $b \in \{0,1\}$
- k significant bits after the initial 1, i = Integer(k.bits)
 - 1. b..b b with $b \in \{0, 1\}$

$$\bullet \ \alpha_i = \frac{2^k}{2^k + i}$$

•
$$m = \frac{m \times \alpha_i}{\alpha_i} \Rightarrow \log(m) = \log(m * \alpha_i) - \log(\alpha_i)$$

•
$$1 < m' < 1 + \frac{1}{2^k}$$
 with $m' = m \times \alpha_i$

1.
$$0..0$$
b with $b \in \{0, 1\}$



•
$$1 < m' < 1 + \frac{1}{2^k}$$

•
$$\log(1+t)$$

•
$$t = m' - 1$$

•
$$0 < t \le \frac{1}{2^k}$$

•
$$1 < m' < 1 + \frac{1}{2^k}$$

•
$$\log(1+t)$$

•
$$t = m' - 1$$

•
$$0 < t \le \frac{1}{2^k}$$

•
$$1 < m' < 1 + \frac{1}{2^k}$$

•
$$log(1+t)$$

•
$$t = m' - 1$$

•
$$0 < t \le \frac{1}{2^k}$$

•
$$1 < m' < 1 + \frac{1}{2^k}$$

•
$$log(1+t)$$

•
$$t = m' - 1$$

•
$$0 < t \le \frac{1}{2^k}$$

Result construction

Operator precision calculation

•
$$X = \circ(x)$$

• relative error =
$$\frac{X-x}{|x|}$$

•
$$X = x.(1 + \epsilon)$$
 with $|\epsilon| \le u$.

Operator precision calculation

•
$$X = \circ(x)$$

• relative error =
$$\frac{X-x}{|x|}$$

•
$$X = x.(1 + \epsilon)$$
 with $|\epsilon| \le u$.

Operator precision calculation

•
$$X = \circ(x)$$

• relative error =
$$\frac{X-x}{|x|}$$

•
$$X = x.(1 + \epsilon)$$
 with $|\epsilon| \le u$.

Example of operator precision calculation

Algorithme

Algorithm: Add122

Input: a Double number and (b_h, b_ℓ) Double-Double number

Condition: $|a| \geq |b_h|$ and $|b_\ell| \leq u.|b_h|$

Output: s and t are Double numbers. s is the main value and t is the error value.

$$2 t = \ell + b_{\ell}$$

$$|a| > |b_h|$$

$$\bullet$$
 $a>0$ and $b_h>0$

$$t = (\ell + b_{\ell}) \cdot (1 + \epsilon) \text{ with } |\epsilon| \le u$$

$$t = \ell + b_{\ell} + \delta$$

with
$$\delta = (\ell + b_\ell).\epsilon$$

$$|\delta| < |\epsilon| \cdot (|\ell| + |b_{\ell}|)$$

•
$$|s+\ell| \ge (1-u).|s| \Rightarrow (1-u).|s| \le |a+b_h| \Rightarrow |s| \le \frac{1}{1-u}.|a+b_h| \Rightarrow |\ell| \le \frac{u}{1-u}.|a+b_h|$$

$$|\delta| \leq |\epsilon| \cdot (\frac{u}{1-u} \cdot |a+b_h| + |b_\ell|)$$



Example of operator precision calculation

Algorithme

•

Algorithm: Add122

Input : a Double number and (b_h, b_ℓ) Double-Double number

Condition: $|a| \geq |b_h|$ and $|b_\ell| \leq u.|b_h|$

Output: s and t are Double numbers. s is the main value and t is the error value.

- 1 return s, t
- $|\delta| \leq |\epsilon| \cdot \frac{-u^2 + 2 \cdot u}{1 u} \cdot |a + b_h|$

$$|\delta| \leq 2.u^2.|a+b_h|$$

$$|a + b_h + b_\ell| \le |a + b_h| + |b_\ell|$$

$$|a + b_h + b_\ell| \ge (1 - u)|a + b_h|$$

$$|\delta| \leq |a + b_h + b_\ell|(\frac{1}{1-u}).2.u^2$$

Example of operator precision calculation

Algorithme

Algorithm: Add122

Input : a Double number and (b_h, b_ℓ) Double-Double number

Condition: $|a| \geq |b_h|$ and $|b_\ell| \leq u.|b_h|$

Output: s and t are Double numbers. s is the main value and t is the error value.

$$2 t = \ell + b_{\ell}$$

$$|\delta| \leq 2.u^2.|a+b_h|$$

$$|a + b_h + b_\ell| \le |a + b_h| + |b_\ell|$$

$$|s+b_h+b_\ell| \geq (1-u)|s+b_h|$$

$$|\delta| \leq |a+b_h+b_\ell|(\frac{1}{1-u}).2.u^2$$

$$|\epsilon| \leq \frac{2.u^2}{1-u} \leq 2.u^2$$

•
$$s + t = (a + (b_h + b_\ell))(1 + \epsilon)$$
 with $|\epsilon| \le 2.u^2$

crlog_{fast}

Algorithme

Algorithm: crlogfast (Part 1)

Condition x is a double number and is not to close to 1.

Condition m_1 and i are results made after calculating x.

Condition $table_{\alpha_i}$ $table_{\log(\alpha_i)}$ and are calculated with our method with Tang and Gal's method.

Condition log(2) in double-double: $h_{log 2} = 0 \times 1.62 e 42 fefa 38 p - 1$ and $h_{log 2} = 0 \times 1.e f 35793 c 7673 p - 45$

Input : x is a double number, i is a Integer number and m_1 is double number. Output : Approximation of log(x) in double number.

1
$$h_r$$
, $\ell_r = Mu/122(m_1, \alpha_i)$

2
$$h, \ell = Add122(-1, h_r, \ell_r)$$

4
$$hh_2$$
, $\ell\ell_2 = Mul222(h, \ell, h, \ell)$

5
$$hh_4$$
, $\ell\ell_4 = Mul222(hh_2, \ell\ell_2, hh_2, \ell\ell_2)$

6
$$ffh_0$$
, $ff\ell_0 = Mul122(f_1, h, \ell)$

crlog_{fast}

Algorithme

Algorithm : crlog_{fast} (Part 2)

 $\mathbf{P} h_{\mathbf{4}}, \ell_{\mathbf{4}} = Add222(ffhx_{\mathbf{0}}, ff\ell x_{\mathbf{0}}, ffhx_{\mathbf{4}}, ff\ell x_{\mathbf{4}})$

13 $h_5, \ell_5 = Add122(\log_{\alpha_i}, h_4, \ell_4)$

 $e_{hlog2}, e_{\ell log2} = Mul122(e, h_{log2}, \ell_{log2})$

15 h_{6} , $\ell_{6} = Add222(e_{hlog2}, e_{\ell log2}, h_{5}, \ell_{5})$

6 return (h_6, ℓ_6)

crlog_{accurate}

Algorithme

Algorithm: crlogaccurate

Condition x is a double number and is not to close to 1.

Condition m_1 and i are results made after calculating x.

Condition $table_{\alpha_i}$ $table_{\log(\alpha_i)}$ are of tables of Double-Double and are calculated with Tang.

Condition log(2) in double-double: $h_{log 2} = 0 \times 1.62 e 42 fefa 39 e fp - 1$ and $h_{log 2} = 0 \times 1.abc 9 e 3b 39803 fp - 56$

Input: \times is a double number, i is a Integer number and m_1 is a double number.

Output : Approximation of log(x) in double number.

2
$$h, \ell = Add122(-1, h_r, \ell_r)$$

3
$$h_1, \ell_1 = Mul222(G[0][0], G[0][1], h, \ell)$$

4 for
$$(j = 1, j < 11, j + +)$$

6
$$h_1, \ell_1 = Mul222(h_2, \ell_2, h, \ell)$$

$$n_{5}, \ell_{5} = Add222(\log_{\alpha_{i}}, h_{1}, \ell_{1})$$

8
$$e_{hlog 2}, e_{\ell log 2} = Mul122(e, h_{log 2}, \ell_{log 2})$$

$$\mathbf{0}$$
 return $h_{\mathbf{6}} + \ell_{\mathbf{6}}$

crlog_{advanced}

Algorithme

Algorithm : crlogadvanced

Input : x is a double number, i is a Integer number. Output : Approximation of log(x) in double number.

(2)
$$h, m, \ell = Add133(-1, h_r, \ell_r)$$

3
$$h_1, m_1, \ell_1 = Mul333(U[0][0], U[0][1], U[0][2], h, m, \ell)$$

4 for
$$(j = 1, j < 14, j + +)$$

5
$$h_2, m_2, \ell_2 = Add333(U[j][0], U[j][1], U[j][2], h_1, m_1, \ell_1)$$

$$h_1, m_1, \ell_1 = Mu/333(h_2, m_2, \ell_2, h, m, \ell)$$

$$h_{5}, m_{5}, \ell_{5} = Add333(\log_{\alpha_{i}} h, \log_{\alpha_{i}} m, \log_{\alpha_{i}} \ell, h_{1}, m_{1}, \ell_{1})$$

8
$$e_{hlog2}$$
, e_{mlog2} , $e_{\ell log2} = Mul133(e, h_{log2}, m_{log2}, \ell_{log2})$

$$m_\ell = m_6 + \ell_6$$

$$mathred m_{\ell}$$
 return $h_{6} + m_{\ell}$

Algorithme

```
Algorithm: log
```

Condition :x is a double number.

Condition : $err_{fast} = 0 \times 1.6b6b11ea279ecp - 59$ and $err_{accurate} = 0 \times 1.810c9a86fc45ep - 99$

Output :log(x) in double number.

2 left =
$$h_6$$
 + FMA($-err_{fast}$, h_6 , ℓ_6)

3
$$right = h_6 + FMA(+err_{fast}, h_6, \ell_6)$$

$$4$$
 if (right == left)

$$6 h_{\mathbf{6}}, l_{\mathbf{6}} = cr \log_{accurate}(x)$$

1 Ieft =
$$h_6$$
 + FMA($-err_{fast}$, h_6 , ℓ_6)

8 right =
$$h_6 + FMA(+err_{accurate}, h_6, \ell_6)$$

$$\bigcirc$$
 if (right == left)

- This code has been verified thanks to Sage and the MPFR library in C.
- ./perf.sh log:
 84.164 :number of cycles of our logarithm
 15.918 : number of cycles of logarithm of system library (GNU libc).
- More information about COREMATH project: https://core-math.gitlabpages.inria.fr

Introduction
Floating point numbers
Steps to calculate crlog
Algorithm
Conclusion

