ε-pTEX の浮動小数点演算の簡易説明書

北川 弘典

2008 年 3 月 16 日

本文章では，ε-pTEX で実装されている浮動小数点演算について概説する．詳細な実装方法についてはソース（`eptex-fp.c`, `fp-mpfr.ch`）に譲ることにして，ここでは簡単に使い方のみ述べる．

本文章でいう**浮動小数点数**とは，よくあるように`-235.673578432E-534` のように，符号と 10 進小数からなる仮数部に，必要に応じて `E` or `e` で始まる指数部が続いたものである．符号は複数あってもよく，そのときは全部掛け算したものが最終的な符号となる．小数点は日米などで使われるピリオドも，欧州大陸で使われるコンマも許容される．

なお，浮動小数点演算を使う場合は，ε-TEX でいうところの extended mode で処理しなければならない．`README.txt` にそってフォーマットファイルを作り，それを使った場合は，extended mode は最初から on になっている．

ε-pTEX 80131.21 版以前では，本文書に書かれている内容は適用されない．

1 80131.21 版以前との変更点

- 演算部を自前で書いていたのを，MPFR library (`http://www.mpfr.org/`) を利用するように変えた．MPFR library は GMP library (`http://gmplib.org/`) を内部で使用するため，両者のインストールが必要である．なお，開発環境は MPFR が 2.3.0，GMP が 4.2.2 である．
- その影響で，格納 format も変わり，10 進 21 桁から 2 進 78 bit (implicit 79 bit)，指数部 16bit となった．扱える数値の範囲は減少したが，有効桁数は 2 桁ほど上がっている．
- `\fpfrac` の出力する桁数が標準で 23 桁となった．桁数は（新設した）`\fpoutprec` パラメータを変更することによって変えられる．
- Overflow, NaN 時のエラー処理は取り除いた．ただ単に現時点では面倒だったから．MPFR 自体が Overflow, Underflow, Inexact, NaN, Range error という 5 種類の例外に対応しているので，これらに対応させることも検討中.
- `\fpinit`, `\fpdest`, `\fppowi` は不要になったので取り除いた.

2 代入，型変換，入出力

■`\real` ⟨*real*⟩ 浮動小数点数を表現する glue（以下，⟨*f-glue*⟩ と称する）を返す.

■`\fpfrac` ⟨*f-glue*⟩ 引数の仮数部を返す.

■`\fpexpr` ⟨*f-glue*⟩ 引数の指数部を返す．上の`\fpexpr` と対にして用いる.

■`\fpoutprec` 上の 2 つのコマンドで出力する桁数を格納するパラメータ．このパラメー

タの値だけ桁が出力される．負数と 30 より大きな値を指定した場合は 23 と同義に扱われる．初期値は 0.

■`\fptoint` ⟨*f-glue*⟩　引数を整数に変換したものを返す．範囲内に収まらないときは，`Number too big` エラーを返す．なお，引数が整数でないときは，0 に近い方に丸められる．

■`\fptodim` ⟨*f-glue*⟩　引数を dimension に，1.0 がちょうど 1 pt になるように変換したものを返す．範囲内に収まらないときは，`Dimension too large` エラーを返す．なお，引数が 1/65536 (1 sp に対応) でないときは，同じように 0 に近い方に丸められる．

3　四則演算等

■`\fpadd` ⟨*f-reg*⟩ ⟨*real*⟩　⟨*f-glue*⟩ が格納された skip レジスタ（以下，⟨*f-reg*⟩ と称する）と浮動小数点数を引数にとり，2 つの浮動小数点数の和を計算して，⟨*f-glue*⟩ に上書きする．

■`\fpsub` ⟨*f-reg*⟩ ⟨*real*⟩　同様に差を計算して，⟨*f-glue*⟩ に上書きする．

■`\fpmul` ⟨*f-reg*⟩ ⟨*real*⟩　同様に積を計算して，⟨*f-glue*⟩ に上書きする．

■`\fpdiv` ⟨*f-reg*⟩ ⟨*real*⟩　同様に商を計算して，⟨*f-glue*⟩ に上書きする．

■`\fppow` ⟨*f-reg*⟩ ⟨*real*⟩　同様に累乗を計算して，⟨*f-glue*⟩ に上書きする．以前の版では $0^0 = 1$ としていたので，（MPFR の実装とは異なるが）継承してある．

4　単項演算

以下は単項演算で，⟨*f-reg*⟩ を 1 つとり，演算をし，結果をその ⟨*f-reg*⟩ に上書きする．よって，以下はコマンドの引数も省略し，演算内容しか書かない．引数の内容は便宜的に x で表す．

■`\fpneg`　-1 倍を計算する．

■`\fpsqr`　平方根 $\sqrt{x}$ を計算する．

■`\fpexp`　指数関数 $\exp x$ を計算する．

■`\fplog`　対数関数 $\log x$ を計算する．

■`\fpabs`　絶対値を計算する．

■`\fpceil`　天井関数 $\lceil x \rceil$，つまり x を越えない最小の整数を計算する．

■`\fpfloor`　床関数 $\lfloor x \rfloor$，つまり x 以下の最大の整数を計算する．

■`\fpsin`, ..., `\fptan`　それぞれ三角関数 $\sin x$, $\cos x$, $\tan x$ を計算する．

■`\fpsinh`, ..., `\fptanh`　それぞれ双曲線関数 $\sinh x$, $\cosh x$, $\tanh x$ を計算する．

■`\fpasin`, ..., `\fpatan`　それぞれ逆三角関数 $\arcsin x$, $\arccos x$, $\arctan x$ を計算する．
結果は $\arcsin x$, $\arctan x \in [-\pi/2, \pi/2]$, $\arccos x \in [0, \pi]$ である（主値をとってる）．

■`\fpasinh`, ..., `\fpatanh`　それぞれ逆双曲線関数 $\operatorname{arcsinh} x$, $\operatorname{arccosh} x$, $\operatorname{arctanh} x$ を計算する．
結果は $\operatorname{arcsinh} x$, $\operatorname{arctanh} x \in \mathbf{R}$, $\operatorname{arccosh} x \in [0, \infty]$ の範囲に収まる．

5　数値積分によるサンプル

本節では，$f(x) = 1/(x+3)$ を $[-1, 1]$ で，区間を 40 等分に分割して台形則，中点則，Simpson 法による数値積分を行う．当然ながら真値は $\log 2 \simeq 6.9314718055994530941723 \times 10^{-1}$ である．

たたき台として，以下の Fortran 90 のプログラムを使用した．これは 2007 年度夏学期の東京大学理学部数学科の講義「計算数理 I」で僕が提出したレポートの中にあったプログラムを簡略化したものである．

```
PROGRAM main
  IMPLICIT REAL*8 (a-h,o-z)
  a=-1d0; b=1d0; n=40; d=0d0; u=0d0
  DO i=0,n-1
    u=u+1d0/(3d0+a+(b-a)*i/n)
    d=d+1d0/(3d0+a+(b-a)*(i+5d-1)/n)
  END DO
  u=u+5d-1*1d0/(3d0+b)-5d-1*1d0/(3d0+a)
  WRITE(*,*) 'DAIKEI: ', u*(b-a)/n
  WRITE(*,*) 'CHUTEN: ', d*(b-a)/n
  WRITE(*,*) 'SIMPSON: ', (u+2d0*d)*(b-a)/n/3d0
  END
```

これの実行結果は以下である．

```
[h7k doc]$ gfortran -o ks1 ks1.f90
[h7k doc]$ ./ks1
 DAIKEI:   0.693186240009141
 CHUTEN:   0.693127651979310
 SIMPSON:   0.693147181322587
```

これを ε-pTEX の浮動小数点演算で書き直して計算させたところ，以下の結果になった（有効桁数は 15 桁に合わせてある）：

台形則での計算結果：　$6.93186240009141 \times 10^{-1}$
中点則での計算結果：　$6.93127651979310 \times 10^{-1}$
Simpson 則での計算結果：　$6.93147181322587 \times 10^{-1}$
真値：　$6.93147180559945 \times 10^{-1}$

本文書のソースを示す．ε-TEX の\numexpr 相当の機能がまだ準備されていないので，ソースは無残な姿である．

```
%#!eplatex fp.tex
\documentclass[a4j,papersize]{jsarticle}
\def\epTeX{$\varepsilon$-\pTeX}\def\eTeX{$\varepsilon$-\TeX}
\def\<#1>{$\langle\hbox{\it #1\/}\rangle$}
\def\.#1{{\tt\char'134 #1}}
\def\listx{\def\makelabel{\selectfont ■}\def\@{\hfill}
\labelwidth=14zw\labelsep1zw\itemindent11zw\leftmargin=4zw}
\def\arcsinh{\mathop{\rm arcsinh}}
\def\arccosh{\mathop{\rm arccosh}}
\def\arctanh{\mathop{\rm arctanh}}
\usepackage{moreverb,bm}
\title{\epTeX の浮動小数点演算の簡易説明書}
\author{北川 弘典}
\date{2008 年 3 月 16 日}
\begin{document}

\maketitle
本文章では，\epTeX で実装されている浮動小数点演算について概説する．詳細な
実装方法についてはソース（{\tt eptex-fp.c}, {\tt fp-mpfr.ch}）に譲
ることにして，ここでは簡単に使い方のみ述べる．

本文章でいう{\gt 浮動小数点数}とは，よくあるように
{\tt -235.673578432E-534}のように，符号と 10 進小数からなる仮数部に，必要に
応じて{\tt E} or {\tt e}で始まる指数部が続いたものである．符号は複数あってもよく，
そのときは全部掛け算したものが最終的な符号となる．小数点は日米などで使わ
れるピリオドも，欧州大陸で使われるコンマも許容される．

なお，浮動小数点演算を使う場合は，\eTeX でいうところの extended mode で処
理しなければならない．{\tt README.txt}にそってフォーマットファイルを作り，
それを使った場合は，extended mode は最初から on になっている．

{\gt\bf $\bm\varepsilon$-\pTeX\ 80131.21 版以前では，本文書に書かれている
内容は適用されない．}
\section{80131.21 版以前との変更点}
\begin{itemize}
\item 演算部を自前で書いていたのを，MPFR library（{\tt
      http://www.mpfr.org/}）を利用するように変えた．MPFR library は GMP
      library（{\tt http://gmplib.org/}）を内部で使用するため，両者のイン
      ストールが必要である．なお，開発環境は MPFR が 2.3.0，GMP が 4.2.2 であ
      る．
\item その影響で，格納 format も変わり，10 進 21 桁から 2 進 78\,bit（implicit
      79\,bit），指数部 16bit となった．扱える数値の範囲は減少したが，有効桁数
      は 2 桁ほど上がっている．
\item \.{fpfrac}の出力する桁数が標準で 23 桁となった．桁数は
      （新設した）\.{fpoutprec}パラメータを変更することによって変えられる．
\item Overflow, NaN 時のエラー処理は取り除いた．ただ単に現時点では面倒だっ
```

たから．MPFR 自体が Overflow, Underflow, Inexact, NaN, Range error という 5 種類の例外に対応しているので，これらに対応させることも検討中．

\item \.{fpinit}, \.{fpdest}, \.{fppowi}は不要になったので取り除いた．

\end{itemize}

\section{代入，型変換，入出力}

\begin{list}{}{\listx}

\item[\.{real}\ \<real>\@] 浮動小数点数を表現する glue（以下，\<f-glue>と称する）を返す．

\item[\.{fpfrac}\ \<f-glue>\@] 引数の仮数部を返す．

\item[\.{fpexpr}\ \<f-glue>\@] 引数の指数部を返す．上の\.{fpexpr}と対にして用いる．

\item[\.{fpoutprec}\@] 上の 2 つのコマンドで出力する桁数を格納するパラメータ．このパラメータの値だけ桁が出力される．負数と 30 より大きな値を指定した場合は 23 と同義に扱われる．初期値は 0．

\item[\.{fptoint}\ \<f-glue>\@] 引数を整数に変換したものを返す．範囲内に収まらないときは，{\tt Number too big}エラーを返す．なお，引数が整数でないときは，0 に近い方に丸められる．

\item[\.{fptodim}\ \<f-glue>\@] 引数を dimension に，1.0がちょうど $1\,$pt になるように変換したものを返す．範囲内に収まらないときは，{\tt Dimension too large}エラーを返す．なお，引数が$1/65536$ (1\,sp に対応) でないときは，同じように 0 に近い方に丸められる．

\end{list}

\section{四則演算等}

\begin{list}{}{\listx}

\item[\.{fpadd}\ \<f-reg>\ \<real>\@]
\<f-glue>が格納された skip レジスタ（以下，\<f-reg>と称する）と浮動小数点数を引数にとり，2 つの浮動小数点数の和を計算して，\<f-glue>に上書きする．

\item[\.{fpsub}\ \<f-reg>\ \<real>\@]
同様に差を計算して，\<f-glue>に上書きする．

\item[\.{fpmul}\ \<f-reg>\ \<real>\@]
同様に積を計算して，\<f-glue>に上書きする．

\item[\.{fpdiv}\ \<f-reg>\ \<real>\@]
同様に商を計算して，\<f-glue>に上書きする．

\item[\.{fppow}\ \<f-reg>\ \<real>\@] 同様に累乗を計算して，\<f-glue>に上書きする．以前の版では$0^0=1$としていたので，(MPFR の実装とは異なるが) 継承してある．

\end{list}

\section{単項演算}

以下は単項演算で，\<f-reg>を 1 つとり，演算をし，結果をその\<f-reg>に上書きする．よって，以下はコマンドの引数も省略し，演算内容しか書かない．引数の内容は便宜的にxで表す．

\begin{list}{}{\listx}

```
\item[\.{fpneg}\@] $-1$倍を計算する．
\item[\.{fpsqr}\@] 平方根$\sqrt x$を計算する．
\item[\.{fpexp}\@] 指数関数$\exp x$を計算する．
\item[\.{fplog}\@] 対数関数$\log x$を計算する．
\item[\.{fpabs}\@] 絶対値を計算する．
\item[\.{fpceil}\@] 天井関数$\lceil x \rceil$，つまり$x$を越えない最小の
                    整数を計算する．
\item[\.{fpfloor}\@] 床関数$\lfloor x \rfloor$，つまり$x$以下の最大の
                    整数を計算する．
\item[\.{fpsin}, $\ldots\,$, \.{fptan}\@]
 それぞれ三角関数$\sin x$, $\cos x$, $\tan x$を計算する．
\item[\.{fpsinh}, $\ldots\,$, \.{fptanh}\@]
 それぞれ双曲線関数$\sinh x$, $\cosh x$, $\tanh x$を計算する．
\item[\.{fpasin}, $\ldots\,$, \.{fpatan}\@] それぞれ逆三角関数$\arcsin
 x$, $\arccos x$, $\arctan x$を計算する．\\ 結果は$\arcsin x$, $\arctan
 x\in [-\pi/2,\pi/2]$, $\arccos x\in [0,\pi]$である（主値をとってる）．
\item[\.{fpasinh}, $\ldots\,$, \.{fpatanh}\@] それぞれ逆双曲線関数
$\arcsinh x$, $\arccosh x$, $\arctanh x$を計算する．\\結果は$\arcsinh x$,
$\arctanh x\in {\bf R}$, $\arccos x\in [0,\infty]$の範囲に収まる．
\end{list}

\section{数値積分によるサンプル}
本節では，$f(x)=1/(x+3)$を$[-1,1]$で，区間を 40 等分に分割して台形則，中
点則，Simpson 法による数値積分を行う．当然ながら真値は
\skip300=\real2\fplog\skip300$\log2\simeq\fpfrac\skip300\times
10^{\fpexpr\skip300}$
である．

たたき台として，以下の Fortran 90 のプログラムを使用した．これは 2007 年度夏
学期の東京大学理学部数学科の講義「計算数理 I」で僕が提出したレポートの中
にあったプログラムを簡略化したものである．
\begin{verbatim}
PROGRAM main
  IMPLICIT REAL*8 (a-h,o-z)
  a=-1d0; b=1d0; n=40; d=0d0; u=0d0
  DO i=0,n-1
    u=u+1d0/(3d0+a+(b-a)*i/n)
    d=d+1d0/(3d0+a+(b-a)*(i+5d-1)/n)
  END DO
  u=u+5d-1*1d0/(3d0+b)-5d-1*1d0/(3d0+a)
  WRITE(*,*) 'DAIKEI: ', u*(b-a)/n
  WRITE(*,*) 'CHUTEN: ', d*(b-a)/n
  WRITE(*,*) 'SIMPSON: ', (u+2d0*d)*(b-a)/n/3d0
  END
\end{verbatim}
これの実行結果は以下である．
\begin{verbatim}
[h7k doc]$ gfortran -o ks1 ks1.f90
[h7k doc]$ ./ks1
```

```
 DAIKEI:   0.693186240009141
 CHUTEN:   0.693127651979310
 SIMPSON:   0.693147181322587
\end{verbatim}

% 以降にプログラムに入る．とりあえずは上のものを逐語訳する方向でいこう．
これを\epTeX の浮動小数点演算で書き直して計算させたところ，以下の結果に
なった（有効桁数は 15 桁に合わせてある）：

\par\vskip0.5\baselineskip\fpoutprec=15\par
\newskip\nia\newskip\nib\newskip\nid
\newskip\niu\newcount\nin\newcount\nii
%
\nia=\real-1 \nib=\real1 \nin=40 \nid=\real0 \niu=\nid % line 3
%
\nii=0
\loop \ifnum\nii<\nin\relax
   \skip300=\real3 \fpadd\skip300\nia % \skip300 = 3+a
   \skip301=\real\nii \fpdiv\skip301\nin % \skip301=i/n
   \skip302=\nib \fpsub\skip302\nia \fpmul\skip302\skip301 % \skip302=(b-a)*i/n
   \skip301=\skip300 \fpadd\skip301\skip302 % \skip301=3+a+(b-a)*i/n
   \fppow\skip301 by -1 \fpadd\niu\skip301 % line 5
   \skip301=\real\nii \fpadd\skip301by0.5 \fpdiv\skip301\nin
   \skip302=\nib \fpsub\skip302\nia \fpmul\skip302\skip301
   \skip301=\skip300 \fpadd\skip301\skip302
   \fppow\skip301 by -1 \fpadd\nid\skip301 % line 6
   \advance\nii by1
\repeat
%
\skip300=\real0.5 \skip301=\real3 \fpadd\skip301\nib
\fpdiv\skip300\skip301
\skip301=\real0.5 \skip302=\real3 \fpadd\skip302\nia
\fpdiv\skip301\skip302
\fpsub\skip300\skip301 \fpadd\niu\skip300 % line 8
%
\skip300=\nib\fpsub\skip300\nia\fpdiv\skip300by\nin
\fpmul\niu\skip300 \fpmul\nid\skip300 % 先に (b-a)/n で掛けておく
%
\noindent
\leavevmode\hbox to 13zw{台形則での計算結果：\hss}%
$\fpfrac\niu\times 10^{\fpexpr\niu}$\\
%
\leavevmode\hbox to 13zw{中点則での計算結果：\hss}%
$\fpfrac\nid\times 10^{\fpexpr\nid}$\\
%
\leavevmode\hbox to 13zw{Simpson 則での計算結果：\hss}%
\skip300=\niu\fpadd\skip300\nid\fpadd\skip300\nid\fpdiv\skip300by3
$\fpfrac\skip300\times 10^{\fpexpr\skip300}$\\
```

```
%
\leavevmode\hbox to 13zw{真値：\hss}%
\skip300=\real2\fplog\skip300
$\fpfrac\skip300\times 10^{\fpexpr\skip300}$

\newpage
本文書のソースを示す．\eTeX の\verb+\numexpr+ 相当の機能がまだ準備されて
いないので，ソースは無残な姿である．
\small
\listinginput[5]{1}{fp.tex}
\end{document}
```