浮動小数点演算プリミティブの実装基準（仮）

October 3, 2023

浮動小数点数のフォーマットは IEEE 単精度フォーマットに基づく。ただし，同等の機能が得られれば内部的に異なる表現を用いても構わない。IEEE 規格の $+0$ とノーマル数が表現でき，かつ，これらの浮動小数点値に対しては対応する実数値が定義されていなければならない。それ以外の浮動小数点値の定義／未定義および値の解釈は指定しないが，個々の浮動小数点値について対応する実数値が定義されているか否かを一意に定めねばならない。以下では，対応する実数値が定義されている浮動小数点数を「有効な浮動小数点数」と呼ぶ。

以下の説明では，$\varepsilon = 2^{-126}$ とする。また，数学的関数と識別できるように，浮動小数点演算プリミティブはキャピタライズしたうえで “FADD” のようにタイプライタ体を用いて表記する。

fadd　有効な浮動小数点数 A, B の組が $-2^{127} < A,\ B,\ A+B < 2^{127}$ を満たすとき，$\mathtt{FADD}(A,\ B)$ は有効な浮動小数点数を返し，以下の条件を満たすこと。

$$|\mathtt{FADD}(A,\ B) - (A+B)| < \max\left(|A|\cdot 2^{-23},\ |B|\cdot 2^{-23},\ |A+B|\cdot 2^{-23},\ \varepsilon\right)$$

fsub　有効な浮動小数点数 A, B の組が $-2^{127} < A,\ B,\ A-B < 2^{127}$ を満たすとき，$\mathtt{FSUB}(A,\ B)$ は有効な浮動小数点数を返し，以下の条件を満たすこと。

$$|\mathtt{FSUB}(A,\ B) - (A-B)| < \max\left(|A|\cdot 2^{-23},\ |B|\cdot 2^{-23},\ |A-B|\cdot 2^{-23},\ \varepsilon\right)$$

fmul　有効な浮動小数点数 A, B の組が $-2^{127} < A,\ B,\ AB < 2^{127}$ を満たすとき，$\mathtt{FMUL}(A,\ B)$ は有効な浮動小数点数を返し，以下を満たすこと。

$$|\mathtt{FMUL}(A,\ B) - (AB)| < \max\left(|AB|\cdot 2^{-22},\ \varepsilon\right)$$

fdiv　有効な浮動小数点数 A, $B (B \neq 0)$ の組が $-2^{127} < A,\ B,\ \frac{A}{B} < 2^{127}$ を満たすとき，$\mathtt{FDIV}(A,\ B)$ は有効な浮動小数点数を返し，以下を満たすこと。

$$\left|\mathtt{FDIV}(A,\ B) - \frac{A}{B}\right| < \max\left(\left|\frac{A}{B}\right|\cdot 2^{-20},\ \varepsilon\right)$$

sqrt　有効な浮動小数点数 A が $0 \leq A < 2^{127}$ を満たすとき，$\mathtt{SQRT}(A)$ は有効な浮動小数点数を返し，以下の条件を満たすこと。

$$\left|\mathtt{SQRT}(A) - \sqrt{A}\right| < \max\left(\sqrt{A}\cdot 2^{-20},\ \varepsilon\right)$$

sin　ある定数 $c\,(1-2^{-23} < c < 1+2^{-23})$ が存在して，$-2^{127} < A < 2^{127}$ を満たす全ての有効な浮動小数点数 A に対して $\mathtt{SIN}(A)$ は有効な浮動小数点数を返し，以下の条件を満たすこと。

$$|\mathtt{SIN}(A) - \sin(cA)| < \max\left(|\sin(cA)|\cdot 2^{-18},\ \varepsilon\right)$$

cos　ある定数 $c\,(1-2^{-23} < c < 1+2^{-23})$ が存在して，$-2^{127} < A < 2^{127}$ を満たす全ての有効な浮動小数点数 A に対して $\texttt{COS}(A)$ は有効な浮動小数点数を返し，以下の条件を満たすこと。

$$|\texttt{COS}(A) - \cos(cA)| < \max\left(|\cos(cA)| \cdot 2^{-18},\, \varepsilon\right)$$

atan　$-2^{127} < A < 2^{127}$ を満たす全ての有効な浮動小数点数 A に対して $\texttt{ATAN}(A)$ は有効な浮動小数点数を返し，以下の条件を満たすこと。

$$|\texttt{ATAN}(A) - \arctan(A)| < \max\left(|\arctan(A)| \cdot 2^{-20},\, \varepsilon\right)$$

fhalf　$\texttt{FHALF}(A)$ は $\texttt{FMUL}(A,\, 0.5)$ が満たすべき基準を満たすこと。

fsqr　$\texttt{FSQR}(A)$ は $\texttt{FMUL}(A,\, A)$ が満たすべき基準を満たすこと。

fabs　$-2^{127} < A < 2^{127}$ を満たす全ての有効な浮動小数点数 A に対して $\texttt{FABS}(A)$ は有効な浮動小数点数を返し，$\texttt{FABS}(A) = |A|$ を満たすこと。

fneg　$-2^{127} < A < 2^{127}$ を満たす全ての有効な浮動小数点数 A に対して $\texttt{FNEG}(A)$ は有効な浮動小数点数を返し，$\texttt{FNEG}(A) = -A$ を満たすこと。

fless　有効な浮動小数点数 A, B の組が $-2^{127} < A,\, B < 2^{127}$ を満たすとき，$\texttt{FLESS}(A,\, B)$ は真偽値を返し，$A < B$ と $\texttt{FLESS}(A) = \texttt{true}$ が同値であること。

fiszero　有効な浮動小数点数 A が $-2^{127} < A < 2^{127}$ を満たすとき，$\texttt{FISZERO}(A)$ は真偽値を返し，$A = 0$ と $\texttt{FISZERO}(A,\, B) = \texttt{true}$ が同値であること。

fispos　$\texttt{FISPOS}(A)$ は $\texttt{FLESS}(0.0,\, A)$ が満たすべき基準を満たすこと。

fisneg　$\texttt{FISNEG}(A)$ は $\texttt{FLESS}(A,\, 0.0)$ が満たすべき基準を満たすこと。

floor　有効な浮動小数点数 A が $-2^{127} < A < 2^{127}$ を満たすとき，$\texttt{FLOOR}(A)$ は有効な浮動小数点数を返し，その値は整数で，かつ，$\texttt{FLOOR}(A) \leq A < \texttt{FLOOR}(A) + 1$ が成立すること。

ftoi (int_of_float)　有効な浮動小数点数 A が $-2^{31}+1 \leq A \leq 2^{31}-1$ を満たすとき，$\texttt{FTOI}(A)$ は 32bit 整数を返し，かつ，$|I - A| < |\texttt{FTOI}(A) - A|$ を満たすような 32bit 整数 I が存在しないこと。

itof (float_of_int)　32bit 整数 I に対して $\texttt{ITOF}(A)$ は有効な浮動小数点数を返し，かつ，$|A - I| < |\texttt{ITOF}(I) - I|$ を満たすような有効な浮動小数点数 A が存在しないこと。