数値計算講義ノート２
浮動小数点数(承前)

大石進一

April 16, 2003

§1 IEEE Standard 754 for Binary Floating-Point Arithmetic

第１回目の講義ではIEEE754という浮動小数点数の規格について学んだ。
これについて、まず、復習してみよう。

IEEE754規格では、

1. 浮動小数点の基本フォーマット,拡張フォーマット

2. 四則演算・平方根・剰余・比較命令

3. 整数・浮動小数点間のフォーマット変換

4. 異なる浮動小数点フォーマット間の型変換

5. 浮動小数点と10 進文字列の変換

6. 非数(NaN)を含めた浮動小数点例外

が規定されていて、Pentium など多くのCPUで採用されている。

浮動小数点数には単精度、倍精度があり、それぞれに、基本と拡張がある。これをまとめると次のようになる。

パラメータ	基本単精度	拡張単精度	倍精度	拡張倍精度
有効数字の桁数p	24	32	以上53	64以上
指数の最大値Emax	＋127	＋1023以上	＋1023	＋16383以上
指数の最小値Emin	－126	－1022以下	－1022	－16382以下
指数のバイアス	＋127	規定しない	＋1023	規定しない
指数のビット長	8	11以上	11	15以上
全体のビット長	32	43以上	64	79以上

浮動小数点数はつぎの形をしている。

$$(-1)^s 2^E (b_0.b_1 b_2 \cdots b_{p-1})$$

但し

- s は 0 または 1
- E は E_{min} から E_{max} の間の任意の整数
- b_i は 0 または 1

である。これらには

- $+\infty$ と $-\infty$ の 2 つの無限大

- 最低 1 種類の Signaling NaN

- 最低 1 種類の Quiet NaN

も含まれている。1 つの数値に対して $2^0(1.0) = 2^1(0.1) = 2^2(0.01)$ のように幾つもの表現形式があるとき、冗長であると呼ばれる。このような冗長性があまり起こらないように表現が決められている（拡張フォーマットにおいてのみ起こる）。

$$\pm 2^{Emin}(0.b_1 b_2 \cdots b_{p-1})$$

で表わされる非ゼロの値を非規格化数と呼ぶ。特定の指数の値は非数 (NaN), 無限大 (∞), ゼロと不正規化数のために予約されている。なお、ゼロには符号の情報が付加されており、0 除算のようなケースにおいてその効力を発揮する。

基本フォーマット

基本単精度と基本倍精度で表現される数は3 つのフィールドから構成されている。

1. 1 ビットの符号s

2. げたばき指数e ＝E ＋bias (E はバイアスされていない指数)

3. 有効数字フィールド (仮数部) $f = .b_1 b_2 \cdots b_{p-1}$

単精度

s, e, f	値
$s = any, e = 255, f \neq 0$	NaN
$s = any, e = 255, f = 0$	$(-1)^s \infty$
$s = any, 0 < e < 255, f = any$	$(-1)^s 2^{e-127}(1.f)$
$s = any, e = 0, f \neq 0$	$(-1)^s 2^{e-126}(0.f)$
$s = any, e = 0, f = 0$	$(-1)^s 0$

倍精度

s, e, f	値
$s = any, e = 2047, f \neq 0$	NaN
$s = any, e = 2047, f = 0$	$(-1)^s \infty$
$s = any, 0 < e < 2047, f = any$	$(-1)^s 2^{e-1023}(1.f)$
$s = any, e = 0, f \neq 0$	$(-1)^s 2^{e-1022}(0.f)$
$s = any, e = 0, f = 0$	$(-1)^s 0$

丸め

丸めとは, まず無限の精度で計算を行い、そして必要ならば、デスティネーションのフォーマットに合わせた形に修正し、同時に不正確の例外を発生させるものである。丸めた結果が常に無限の精度で計算したものと同じになることが保証されるように数ビットの保護ビットを用いてこれに対応する。

最近値への丸め

有向丸め

最近値への丸めモ-ドだけでなく、有向丸めモードもユ-ザに選択権を与えなければならない(shall)。有向丸めモ-ドには、＋∞方向への丸め(RP :Round to ＋∞)、−∞方向への丸め(RM :Round to −∞)、0 方向への丸め(RZ :Round to Zero)の3 つのモードがある。

(注意) 後に学ぶが、RP とRM の利用価値は、上限と下限の両方を保証して結果を返す演算(区間演算)にある。区間演算を用いると、誤差の範囲を正確に保証することができる。

命令

この規格では、以下の命令を備えなければならない(shall)。

- 加減乗除
- 平方根
- 剰余
- 浮動小数点の整数への丸め
- 異なる浮動小数点フォ-マットへの変換

- 浮動小数点フォ-マットと整数フォ-マット間の変換
- ２ 進10 進変換
- 比較

これらの命令は、まず無限の精度で計算を行い、その後デスティネ-ションのフォ-マットに合わせて修正を行わうことで定義される。

§2 浮動小数点数をさわってみよう

数値計算ツールを利用しよう。OSとしてLinuxまたはWindows+Cygwin
の利用を考える。

`http://www.cygwin.com/xfree/`

からCygwinをダウンロードしてくる。最新のものでは、Xサーバーもつい
てくる。

ここで、著者(大石進一)の作った数値計算ツールSlab(エスラボと呼んでほ
しい)を使おう。

`http://www.oishi.info.waseda.ac.jp/~oishi/lec2003/Slab121_cyg.tgz`

をダウンロードして使おう。