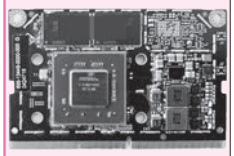


第9章　数値の表し方から足し算，掛け算，丸め方まで

GPUのアーキテクチャ研究⑤ 浮動小数点演算の基礎

浮動小数点データとその演算方法の基礎について説明します．GPUでは浮動小数点データを多く扱うので，その基本を知っておくことには損はないと思います．

9-1　浮動小数点数の表し方

● 浮動小数点データのフォーマット

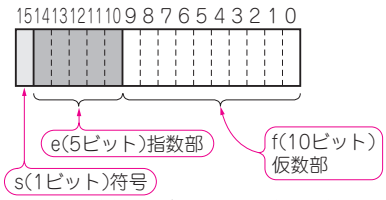
浮動小数点データはその名のごとく，数値を

$$\pm 仮数 \times 2^{指数}$$

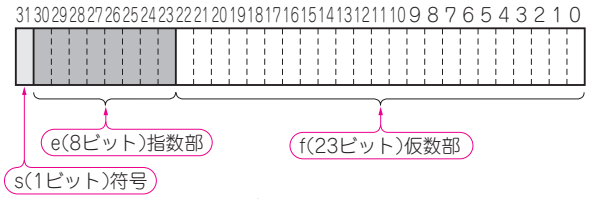
という形式で表現したものです．

ここではGPUでよく使われる，16ビット半精度浮動小数点データ(FP16)，32ビット単精度浮動小数点データ(FP32)，64ビット倍精度浮動小数点データ(FP64)について，いずれも，IEEE754規格に準拠したフォーマットで説明します．もっとビット幅が長いフォーマットもありますが，ここでは省略します．

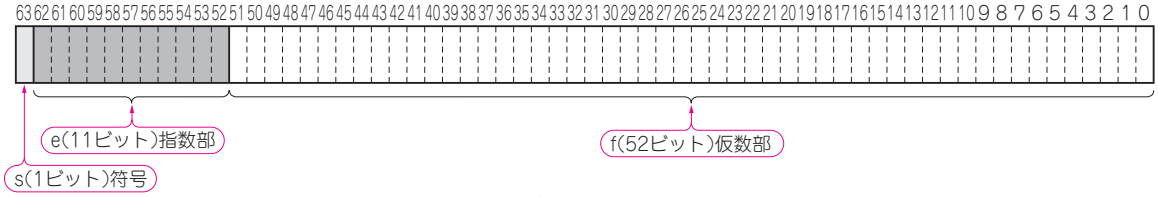
図1にそれぞれのフォーマットを，**表1**にそれぞれ



(a) 16ビット半精度(FP16)



(b) 32ビット単精度(FP32)



(c) 64ビット倍精度(FP64)

図1　浮動小数点データのフォーマット

表1　浮動小数点データの数値表現方法

フォーマット種類		フォーマット			数値表現方法						
		符号部s	指数部e	仮数部f	指数部eの範囲	指数部バイアスBIAS	仮数部 $f=0$ の場合	仮数部 $f\neq 0$ の場合	表現式	表現可能最小値(>0)	表現可能最大値
FP16	16ビット半精度	1ビット	5ビット	10ビット	0x00	14	0または-0	非正規化数	$[(-1)s]\times[2^{-14}]\times[0.f]$	$\pm 5.96\times10^{-8}$	−
					0x01～0x1E	15	正規化数		$[(-1)s]\times[2^{(e-15)}]\times[1.f]$	$\pm 6.10\times10^{-5}$	± 65504
					0x1F	−	± 無限大	非数(NaN)	−	−	−
FP32	32ビット単精度	1ビット	8ビット	23ビット	0x00	126	0または-0	非正規化数	$[(-1)s]\times[2^{-126}]\times[0.f]$	$\pm 1.40\times10^{-45}$	−
					0x01～0xFE	127	正規化数		$[(-1)s]\times[2^{(e-127)}]\times[1.f]$	$\pm 1.18\times10^{-38}$	$\pm 3.40\times10^{38}$
					0xFF	−	± 無限大	非数(NaN)	−	−	−
FP64	64ビット倍精度	1ビット	11ビット	52ビット	0x000	1022	0または-0	非正規化数	$[(-1)s]\times[2^{-1022}]\times[0.f]$	$\pm 4.94\times10^{-324}$	-
					0x001～0x7FE	1023	正規化数		$[(-1)s]\times[2^{(e-1023)}]\times[1.f]$	$\pm 2.23\times10^{-308}$	$\pm 1.80\times10^{308}$
					0x7FF	-	± 無限大	非数(NaN)	-	-	-