【7】

(i) ==========

$(8296)_{10} = 8{,}192 + 64 + 32 + 8 = 2^{13} + 2^6 + 2^5 + 2^3 = (10\ 0000\ 0110\ 1000)_2$ であるから,

$-(8296)_{10} = (-1)^1 \cdot (10\ 0000\ 0110\ 1000)_2 = (-1)^1 \cdot (0.1000\ 0001\ 101)_2 \cdot 2^{14}$ である.

それゆえ, 最上位ビットは 1, 下 1 バイトは 1000 0001 であり, 指数部は, $(14)_{10} = (1110)_2$ にバイアス $(100\ 0000)_2$ を加えて, $(14)_{10} = (100\ 1110)_2^{\text{bias64}}$ となる.

従って, 2 進浮動小数点数の 2 バイトは, (1 100 1110 1000 0001) となる.

この場合, $32 + 8 = 2^5 + 2^3 = (10\ 1000)_2$ は丸められてしまい, この情報は失われる.

(ii) ==========

$(1\ 100\ 1110\ 1000\ 0001)_2$ より, 直ちに $(CE81)_{16}$ と書ける.

(iii) ==========

最大数は 2 バイトが (0 111 1111 1111 1111) となっている場合である.

この指数部は, $(111\ 1111)_2^{\text{bias64}} = (011\ 1111)_2$ であるから,

$(011\ 1111)_2 = (100\ 0000)_2 - 1 = 64 - 1 = 63$ である.

それゆえ, 最大数は, $(-1)^0 \cdot (0.1111\ 1111)_2 \cdot 2^{63}$ である. これは, 下記の数である.

$(0.1111\ 1111)_2 \cdot 2^{63} = (1111\ 1111)_2 \cdot 2^{55} = \{(1\ 0000\ 0000)_2 - 1\} \cdot 2^{55} = \{2^8 - 1\} \cdot 2^{55} = (256-1) \cdot 2^{55}$

$= 255 \times 2^{55}$

(iv) ==========

正の最小数は 2 バイトが (0 000 0000 1000 000) となっている場合である.

この指数部は, $(000\ 0000)_2^{\text{bias64}} = -64$ であるから, 最小数は, 下記の数である.

$(-1)^0 \cdot (0.100\ 0000)_2 \cdot 2^{-64} = 2^{-1} \cdot 2^{-64} = 2^{-65}$

(v) ==========

x の指数部は, $(000\ 1011)_2^{\text{bias64}} = (000\ 1011)_2 - 64 = 11 - 64 = -53$ であり,

y の指数部は, $(100\ 1101)_2^{\text{bias64}} = (100\ 1101)_2 - 64 = (000\ 1101)_2 = 13$ である.

従って, x および y はそれぞれ次のような数であることが分かる.

$$x = (-1)^1 \cdot (0.1010\ 1001)_2 \cdot 2^{-53} \quad , \quad y = (-1)^0 \cdot (0.1001\ 1100)_2 \cdot 2^{13}$$

従って, 2 進 8 桁の有効数字の加算をした場合には, x は y に対して無視できることが分かるから, x＋y は y と同じであり, (0 100 1101 1001 1100) となる.

このように, 浮動小数点数(float や double と宣言された変数)の加減算において, 値が無視されるような数 x_i ($i = 1,2,\cdots$)を何回 y に加減算しても y は変化しない. 従って, プログラムでは, 先に x_i の総和 $\sum_i x_i$ を計算し, それと y の演算を行う必要がある.