Floating-point Examples (SAS on PC and RS6000 AIX)

Introductory Example

A floating-point number consists of at least 3 bytes of 1’s and 0’s. The information stored in these bytes includes the sign, exponent, and mantissa of the number. Up to five additional bytes may be used to store the mantissa. SAS, by default, stores numbers in eight bytes.

Numbers are represented in base-2 exponential notation. For example, 13_{10} (read thirteen, base 10) is represented in base-2 as 1101 = 1.101 \times 10^{11}. [The base-10 equivalent of “10^{11}” is 2^{3}.] Here the sign is positive, the exponent is 11_{2}, and the mantissa is 1.101.

The sign of a floating-point number is indicated by the leading bit of the first byte: 0=positive, 1=negative. For the example above, the number is positive.

The remaining 7 bits of the first byte and the first 4 bits of the second byte are for the exponent. In order to allow for negative exponents, a bias is added to the true exponent before it is stored. The bias is 1023_{10} or 11111111_{2}. For the example above, the exponent is 11_{2}. Adding the bias, the stored value is 10000000010_{2}.

The mantissa will always have a leading one to the left of the decimal. The leading one and the decimal are not stored. In the example above, the mantissa is 1.101. Thus the stored value is 101. Any remaining bits are padded with zeroes.

Usually the floating-point number is written in hexadecimal form (base-16). Each group of four bits above may be represented with a single hexadecimal number. For the example, the floating-point number is 402A00_{16}. 

Example 2

What is the floating-point representation of $-483.137_{10}$? First we must convert this number to base-2.

Remember that any integer, $N$, in base-10 can be represented as the sum $\{\ldots + a_3 \times 2^3 + a_2 \times 2^2 + a_1 \times 2^1 + a_0 \times 2^0\}$, where the $a_i$ values are either 0 or 1. If we divide $N$ by 2, and there is a positive remainder, then $a_0=1$. Otherwise, $a_0=0$. Dividing the integer part of $N/2$ by 2, we can determine the value of $a_1$, and so on. The base-2 representation of $N$ is thus $\{\ldots a_3 a_2 a_1 a_0\}$. As shown below, $483_{10} = 111100011_2$.

\[
\begin{array}{c|c|c|c|c}
483/2 & = & 241 + & 1 & a_0 \\
241/2 & = & 120 + & 1 & a_1 \\
120/2 & = & 60 + & 0 & a_2 \\
60/2 & = & 30 + & 0 & a_3 \\
30/2 & = & 15 + & 0 & a_4 \\
15/2 & = & 7 + & 1 & a_5 \\
7/2 & = & 3 + & 1 & a_6 \\
3/2 & = & 1 + & 1 & a_7 \\
\frac{1}{2} & = & 0 + & 1 & a_8 \\
\end{array}
\]

A decimal number between 0 and $1$, $F$, in base-10 can be represented as the sum $\{b_1 \times 2^{-1} + b_2 \times 2^{-2} + b_3 \times 2^{-3} + \ldots\}$, where the $b_i$ values are either 0 or 1. If we multiply $F$ by 2, and the result is greater or equal to one, then $b_1=1$. Otherwise, $b_1=0$. Multiplying the fractional part of $F \times 2$ by 2, we can determine the value of $b_2$, and so on. The base-2 representation of $F$ is thus $\{b_1 b_2 b_3 \ldots\}$. As shown below, $0.137_{10} = .00100011$, truncated to 8-places.

\[
\begin{array}{c|c|c|c|c}
0.137 \times 2 & = & 0.274 + & 0 & b_1 \\
0.274 \times 2 & = & 0.548 + & 0 & b_2 \\
0.548 \times 2 & = & 0.096 + & 1 & b_3 \\
0.096 \times 2 & = & 0.192 + & 0 & b_4 \\
0.192 \times 2 & = & 0.384 + & 0 & b_5 \\
0.384 \times 2 & = & 0.768 + & 0 & b_6 \\
0.768 \times 2 & = & 0.536 + & 1 & b_7 \\
0.536 \times 2 & = & 0.072 + & 1 & b_8 \\
\end{array}
\]

Thus we have that $483.137_{10} \approx 111100011.00100011_2$. In base-2 scientific notation, the approximate value is $1.1110001110010000111 \times 2^{1000}$. Adding the bias to the exponent yields $10000000111$. Taking in account the negative sign and dropping the leading 1 from the mantissa, the 3-byte floating-point representation of $-483.137_{10}$ is

\[
\begin{array}{cccccc}
1 & 1 & 0 & 0 & 0 & 0 \\
\text{byte } & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 0 \\
\text{byte } & 2 & 1 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 \\
\text{byte } & 3 & 0 & 0 & 0 & 0 \\
\end{array}
\]

Note that not all of the mantissa can be stored. The remainder is truncated. If four or more bytes were available, more of the mantissa could be kept. In hexadecimal form, this number is C07E32_{16}. The actual base-10 number represented by this floating-point number is $-483.125_{10}$. 