Michalis Papadopoullos
03114702
mechatr0nic.cy[@]gmail[dot]com

Quad-precision **(according to IEEE-754 standard)**

MIPS is a 32-bit RISC architecture, and according to its ISA it has a total of 32 floating-point registers of 32-bits long each. We would like to implement an arbitrary floating-point data type, of 128-bit length, to extend the precision and range of representable numbers.

We will use a total of 4 32-bit registers to form our 128-bit data type. Using the IEEE-754 standard, our data type will look like the following diagram:

Reg0 (32bit)	Sign (1bit)	Exponent (15bit)	Fraction #3 (16bit)
Reg1 (32bit)	Fraction #2 (32bit)		
Reg2 (32bit)	Fraction #1 (32bit)		
Reg3 (32bit)	Fraction #0 (32bit)		

where, the ***fraction*** part consists of (3x32 + 16) 112 bits, 15 bits for the ***exponent*** part and 1 bit to denote the ***sign*** of the representable number.

The IEEE-754 standard, divides floating point numbers into three parts, as shown above:

1. Sign
2. Exponent
3. Fraction

Placing the ***exponent*** part before ***fraction*** simplifies the sorting of floating-point numbers using integer comparison instructions - as each of the three parts form an integer value.

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(Image: Computer Organization and Design - Representation of 1.0 in 32bit format)

A challenge appeared when numbers with negative exponents were written in the above format. Using the 2's complement notation, negative numbers have a leading '1' bit (MSB) which denote that the number is of a negative sign.

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(Using the previous format, this is the representation of 0.5 in 32bit format)

The desirable notation must therefore represent the most negative exponent as $00000000000000000000000000000000_2$ and the most positive as $11111111111111111111111111111111_2$ to simplify comparisons.

Michalis Papadopoullos
03114702
mechatr0nic.cy[@]gmail[dot]com

This convention is called ***biased-notation***, with the ***bias*** being the number that is subtracted from the normal, unsigned representation to determine the real value of the exponent part.
IEEE-754 uses a *bias* of $2^{(\text{number of bits in exponent}) - 1} - 1$ (*)
therefore, in our quad-precision data type *bias* = 16383.

Quad-precision floating point - according to IEEE-754 standard:

Sign	Exponent (15bits)	Mantissa (112bits)	Meaning
0	000000000000000_2	$0000000...0000000_2$	+0
1	000000000000000_2	$0000000...0000000_2$	-0
0	000000000000000_2	Non-Zero	+Denormal
1	000000000000000_2	Non-Zero	-Denormal
0	111111111111111_2	$0000000...0000000_2$	+Infinity
1	111111111111111_2	$0000000...0000000_2$	-Infinity
0	111111111111111_2	Non-Zero	NaN
1	111111111111111_2	Non-Zero	NaN

A number X in the above (normalized) representation can be calculated as:
$X = (-1)^{\text{sign}} \times (1 + \text{significand}) \times 2^{\text{exponent - bias}}$.
In the normalized form there are no leading ‘0’ in the fraction part, therefore any number in normalized form has a leading ‘1’ in the integer part of the fraction part and is called *hidden bit* and fraction is computed implicitly by adding 1.0 to the significand. Denormalized floating-point numbers, fill up the gap between 0 and the smallest normalized number, allowing us to extend the system’s representable range.

Minimal value (normalized):	1.0×2^{-16382}
Maximal value (normalized):	$1.1111...1 \times 2^{+16383} \approx 2^{+16384}$
Minimal value (denormalized):	$0.0000...1 \times 2^{-16382} = 1.0 \times 2^{-16495}$
Accuracy:	$1.0 \times 2^{-16382} - 1.0000...1 \times 2^{-16382} = 1.0 \times 2^{-16495}$

References:

[1]. Computer Organization & Design: The hardware/software interface
[2]. IEEE Standard 754 for Binary Floating-Point Arithmetic (W. Kahan)

(*): The exponents $00000000000000000000000000000000_2$ and $11111111111111111111111111111111_2$ are reserved for special purpose signals.