Normalized Floating Point Representation

An Example

David Semeraro

A floating point base 2 number

- Consider the number…

$$x = \pm(0.b_1 b_2 b_3) \times 2^{\pm k}$$

$$(k, b_i \in [0,1])$$

The nonnegative Numbers

$0.000 \times 2^{-1} = 0$	$0.000 \times 2^{0} = 0$	$0.000 \times 2^{1} = 0$
$0.001 \times 2^{-1} = 1/16$	$0.001 \times 2^{0} = 1/8$	$0.001 \times 2^{1} = 1/4$
$0.010 \times 2^{-1} = 2/16$	$0.010 \times 2^{0} = 2/8$	$0.010 \times 2^{1} = 2/4$
$0.011 \times 2^{-1} = 3/16$	$0.011 \times 2^{0} = 3/8$	$0.011 \times 2^{1} = 3/4$
$0.100 \times 2^{-1} = 4/16$	$0.100 \times 2^{0} = 4/8$	$0.100 \times 2^{1} = 4/4$
$0.101 \times 2^{-1} = 5/16$	$0.101 \times 2^{0} = 5/8$	$0.101 \times 2^{1} = 5/4$
$0.110 \times 2^{-1} = 6/16$	$0.110 \times 2^{0} = 6/8$	$0.110 \times 2^{1} = 6/4$
$0.111 \times 2^{-1} = 7/16$	$0.111 \times 2^{0} = 7/8$	$0.111 \times 2^{1} = 7/4$

16 Distinct Numbers (24 representations)

Normalization

$$x = \pm(0.1b_2b_3) \times 2^{\pm k}$$

- $b_1 = 1 \rightarrow$ we only need 2 bits for the mantissa.
- We can use a machine word that only has 2 bits in the mantissa to represent normalized floating point numbers of this form.

Normalization

$0.000 \times 2^{-1} = 0$	$0.000 \times 2^{0} = 0$	$0.000 \times 2^{1} = 0$
$0.001 \times 2^{-1} = 1/16$	$0.001 \times 2^{0} = 1/8$	$0.001 \times 2^{1} = 1/4$
$0.010 \times 2^{-1} = 2/16$	$0.010 \times 2^{0} = 2/8$	$0.010 \times 2^{1} = 2/4$
$0.011 \times 2^{-1} = 3/16$	$0.011 \times 2^{0} = 3/8$	$0.011 \times 2^{1} = 3/4$
$0.100 \times 2^{-1} = 4/16$	$0.100 \times 2^{0} = 4/8$	$0.100 \times 2^{1} = 4/4$
$0.101 \times 2^{-1} = 5/16$	$0.101 \times 2^{0} = 5/8$	$0.101 \times 2^{1} = 5/4$
$0.110 \times 2^{-1} = 6/16$	$0.110 \times 2^{0} = 6/8$	$0.110 \times 2^{1} = 6/4$
$0.111 \times 2^{-1} = 7/16$	$0.111 \times 2^{0} = 7/8$	$0.111 \times 2^{1} = 7/4$

Number in green are denormal. They can not be written in normalized form due to restrictions on the exponent ($k \in [0,1]$).

Normalization

$$0.001 \times 2^{-1} \rightarrow 0.100 \times 2^{-3}$$

- Normalizing $1/16$ would require the exponent to be out of range.
- Denormal numbers are numbers that are representable given the number of bits in the mantissa but can not be normalized without causing the exponent to go out of range.