PSU CS Tutors Floating Point Cheat Sheet

Shawn Roberts October 18, 2018

Problem

General Form of a Floating point number:

[ignore bits][sign bit][exponent bits(e_{bin})][fraction bits(f_{bin})]

f_{bin} = # from fraction binary number $\quad$ f_{size} = # number of bits in fraction binary number
e_{bin} = # from exponent binary number $\quad$ k = # number of bits exponent binary number
$f = f_{bin}/2^{f_{size}}$
bias $= 2^{k-1} - 1$; is found based on normalizing the number sets to straddle 0.
$V = (-1)^{\text{sign}} * M * 2^E$

Normalized

$M = 1 + f \qquad E = e_{bin} - bias$

Denormalized

$M = f \qquad E = 1 - bias$

Cases

Case 1:: special cases :: +- inf, NAN

→ If Exp is all 1's and frac all 0's

[sign] inf

→ If Exp is all 1's and frac !all 0's

Nan

Case 2:: small numbers close to 0 :: Denormalized

→ if Exponent is all 0's

$$M = f$$
$$E = 1 - bias$$

Case 3:: large numbers :: Normalized

→ if Exponent if mix

$$M = 1 + f$$
$$E = e_{bin} - bias$$

Alternate equation

$$V = (-1)^{\text{sign}} * (M)2^{E-f_{size}}$$

$f = (f_{bin})$ $\qquad M = (0|2^{f_{size}} + f)$ $\qquad E = (1|e_{bin}) - bias$

explanation

$V = (-1)^{\text{sign}} * M * 2^E$

$$
\begin{aligned}
V &= (-1)^{\text{sign}} * M * 2^E; \leftarrow M, E \\
V &= (-1)^{\text{sign}} * (0|1 + f) * 2^{(1|e_{bin})-bias}; \leftarrow f \\
V &= (-1)^{\text{sign}} * (0|1 + \frac{f_{bin}}{2^{f_{size}}}) * 2^{(1|e_{bin})-bias} \\
V &= (-1)^{\text{sign}} * (0|\frac{2^{f_{size}}}{2^{f_{size}}} + \frac{f_{bin}}{2^{f_{size}}})) * 2^{(1|e_{bin})-bias} \\
V &= (-1)^{\text{sign}} * (0|2^{f_{size}} + f_{bin})\frac{1}{2^{f_{size}}} * 2^{(1|e_{bin})-bias} \\
V &= (-1)^{\text{sign}} * (0|2^{f_{size}} + f_{bin})\frac{2^{(1|e_{bin})-bias}}{2^{f_{size}}} \\
V &= (-1)^{\text{sign}} * (0|2^{f_{size}} + f_{bin})2^{(1|e_{bin})-bias-f_{size}} \\
V &= (-1)^{\text{sign}} * (0|2^{f_{size}} + f_{bin}) * 2^{(1|e_{bin})-bias-f_{size}}
\end{aligned}
$$

If we define $f = (f_{bin})$, $M = (0|2^{f_{size}} + f_{bin})$, and $E = (1|e_{bin}) - bias$ then this becomes much more manageable.

$$V = (-1)^{\text{sign}} * (M)2^{E-f_{size}}$$

Cases for simplified

Case 1:: special cases :: +- inf, NAN

→ If Exp is all 1's and frac all 0's

[sign] inf

→ If Exp is all 1's and frac !all 0's

Nan

Case 2:: small numbers close to 0 :: Denormalized

→ If Exponent is all 0's

$$
\begin{aligned}
M &= f_{bin} \\
E &= 1 - bias
\end{aligned}
$$

Case 3:: large numbers :: Normalized

→ if Exponent if mix

$$
\begin{aligned}
M &= 1 + f_{bin} \\
E &= e_{bin} - bias
\end{aligned}
$$