Representación flotante o Punto flotante

Floating representation or floating point

Ronald Marín Cardona
Ingeniería de sistemas y computación, UTP, Pereira, Colombia
Correo-e: Ronald.marin@utp.edu.co

***Resumen*— Este documento contiene un resumen sobre la representación flotante, tal y como se da tratamiento en la materia Introducción a la Informática. El objetivo es realizar una revisión de la representación flotante, su sintaxis, operado en un sistema de base diez y utilizando los números binarios.**

***Palabras clave*— Numero, Representación flotante, Decimal, Binario.**

***Abstract*— This document contains a summary on the floating representation, as it is treated in the subject Introduction to Computer Science. The goal is to conduct a review of the floating representation, its syntax, and simple mathematical operations, operated on a base ten system and using binary numbers.**

Key Word **— Number, Floating Representation, Decimal, Binary.**

I. INTRODUCCIÓN

Los computadores no pueden representar los números reales o los números complejos de manera exacta, por ello surge el Punto flotante o Representación flotante, pues en muchas ocasiones se mueven rangos de números muy grandes.

Las computadoras aproximan los Número reales mediante el sistema de **números de punto flotante,** utilizando un número finito de Bits para la representación. [1]

II. CONTENIDO

Cómo funcionan los números de punto flotante

El número se descompone en dos partes:

1) Mantisa: (también llamada coeficiente o significando) que contiene los dígitos del número. Mantisas negativas representan números negativos.

2) Exponente: indica dónde se coloca el punto decimal (o binario) en relación al inicio de la mantisa.

3) Signo: indica el signo del número (0= positivo, 1=negativo)

Este formato cumple todos los requisitos:

- Puede representar números de órdenes de magnitud enormemente dispares (limitado por la longitud del exponente).
- Proporciona la misma precisión relativa para todos los órdenes (limitado por la longitud de la mantisa).
- Permite cálculos entre magnitudes: multiplicar un número muy grande y uno muy pequeño conserva la precisión de ambos en el resultado.

Los números de coma flotante decimales normalmente se expresan en **notación científica** $(n, \ldots x10^n)$ con un punto explícito siempre entre el primer y el segundo dígitos. El exponente o bien se escribe explícitamente incluyendo la base, o se usa una **e** para separarlo de la mantisa.

Mantisa	Exponente	Notación científica	Valor en punto fijo
1.5	4	$1.5 \cdot 10^4$	15000
-2.001	2	$-2.001 \cdot 10^2$	-200.1
5	-3	$5 \cdot 10^{-3}$	0.005
6.667	-11	6.667e-11	0.0000000000667

El estándar

Casi todo el hardware y lenguajes de programación utilizan números de punto flotante en los mismos formatos binarios, que están definidos en el estándar **IEEE 754.** Los formatos más comunes son de 32 o 64 bits de longitud total:

Hay algunas peculiaridades:

Formato	Bits totales	Bits significativos	Bits del exponente	Número más pequeño	Número más grande
Precisión sencilla	32	23 + 1 signo	8	$\sim 1.2 \cdot 10^{-38}$	$\sim 3.4 \cdot 10^{38}$
Precisión doble	64	52 + 1 signo	11	$\sim 5.0 \cdot 10^{-324}$	$\sim 1.8 \cdot 10^{308}$

La secuencia de bits es primero el bit del signo, seguido del exponente y finalmente los bits significativos.

El exponente no tiene signo; en su lugar se le resta un desplazamiento (127 para sencilla y 1023 para doble precisión). Esto, junto con la secuencia de bits, permite que los números de

punto flotante se puedan comparar y ordenar correctamente incluso cuando se interpretan como enteros.

Se asume que el bit más significativo de la mantisa es 1 y se omite, excepto para casos especiales.

Hay valores diferentes para cero positivo y cero negativos. Estos difieren en el bit del signo, mientras que todos los demás son 0. Deben ser considerados iguales aunque sus secuencias de bits sean diferentes.

Hay valores especiales no numéricos (NaN, «not a number» en inglés) en los que el exponente es todo unos y la mantisa no es todo ceros. Estos valores representan el resultado de algunas operaciones indefinidas (como multiplicar 0 por infinito, operaciones que involucren NaN, o casos específicos). Incluso valores NaN con idéntica secuencia de bits no deben ser considerados iguales.[2]

Representacion de Números de Punto Flotante [FloatP]:

En general, un número de punto flotante será representado como:

$$\pm d, dd \ldots d \times \beta^e$$

donde *d,dd…d* es denominado el significando, mantisa o fracción que tiene p dígitos, y e es el exponente .

IEEE COMA FLOTANTE:

El estándar del IEEE para aritmética en coma flotante (IEEE 754) es la norma o estándar técnico para computación en coma flotante, establecida en 1985 por el **Instituto de Ingenieros Eléctricos y Electrónicos** (IEEE). La norma abordó muchos problemas encontrados en las diversas implementaciones de coma flotante que las hacían difíciles de usar de forma fiable y portátil. Muchas unidades de coma flotante de hardware utilizan ahora el estándar IEEE 754.

El estándar define:

- Formatos aritméticos: conjuntos de datos de coma flotante binarios y decimales, que consisten en números finitos, incluidos los ceros con signo y los números desnormalizados o subnormales, infinitos y valores especiales "no numéricos" (NaN).
- Formatos de intercambio: codificaciones (cadenas de bits) que se pueden utilizar para intercambiar datos de coma flotante de forma eficiente y compacta.
- Reglas de redondeo: propiedades que deben satisfacerse al redondear los números durante las operaciones aritméticas y las conversiones.
- Operaciones: operaciones aritméticas y otras (como funciones trigonométricas) en formatos aritméticos.
- Manejo de excepciones: indicaciones de condiciones excepcionales, tales como división por cero, desbordamiento, etc.

Desarrollo del estándar:

La versión actual del estándar, denominada **IEEE 754-2008**, publicada en agosto de 2008, se deriva de la versión anterior **(IEEE 754-1985).**

Los formatos binarios en el estándar original se incluyen en el nuevo estándar junto con tres nuevos formatos básicos (uno binario y dos decimales). Para cumplir con el estándar actual, debe ser implementado al menos uno de los formatos básicos tanto como formato aritmético, como de intercambio.

Formatos:

- Números finitos, que pueden ser de base 2 (binario) o de base 10 (decimal). Cada número finito se describe por tres enteros: s (un bit de signo), c (un significando, mantisa o coeficiente) y q (exponente). El valor numérico v de un número finito es:

$$v = (-1)^s x\ c\ x\ b^q$$

- Dos infinitos: - ∞ y ∞.
- Dos tipos de valores no numéricos (NaN): un NaN silencioso (qNaN) y un NaN de señalización (sNaN). Un NaN puede llevar una carga útil que está destinada a la información de diagnóstico que indica la fuente de la NaN. El signo de un NaN no tiene sentido, pero puede ser predecible en algunas circunstancias.
- c debe ser un número entero en el intervalo cerrado $[0, b^p - 1]$. Por ejemplo, si b = 10 y p = 7 entonces c está en el intervalo cerrado [0, 9999999]).

Representación y codificación en memoria

Algunos números pueden tener varias representaciones en el modelo que acaba de ser descrito. Por ejemplo, si b = 10 y p = 7, entonces el número -12,345 puede representarse como -12345 × 10-3, -123450 × 10-4 y -1234500 × 10-5. Sin embargo, para la mayoría de las operaciones, como las operaciones aritméticas, el resultado (valor) no depende de la representación de las entradas.

Para los formatos binarios, la representación se hace única eligiendo el exponente representable más pequeño. Para los números con un exponente en el rango normal (no todos unos o todos ceros), el bit inicial del significando siempre será 1. En consecuencia, el bit 1 principal puede ser implícito en lugar de estar explícitamente presente en la codificación de la memoria. Esta regla se denomina convención de bit principal, o también convención de bits implícita o convención de bits ocultos. La regla permite que el formato de memoria tenga un poco más de precisión. La convención de bit principal no se utiliza para los números subnormales ya que tienen un exponente fuera del rango del exponente normal.

[3]

III. CONCLUSIONES

A continuación se resumirá la siguiente información: la representación flotante es usada por la computadora para resumir y guardar dígitos, ya sean de pequeña o gran magnitud. La representación flotante puede se representada en decimal o en base dos (binario), teniendo en cuenta que su estándar es el IEEE-754 la cual abordó muchos problemas encontrados en las diversas implementaciones de coma flotante que las hacían difíciles de usar de forma fiable.

REFERENCIAS

[1] https://www.inf.utfsm.cl/~parce/cc1/clase18-RP.html
[2] http://puntoflotante.org/formats/fp/
[3] https://es.wikipedia.org/wiki/IEEE_coma_flotante