2020

NÚMEROS DE PUNTO FLOTANTE: Introducción a la Informática

GIAN FRANCO POSSO GIRALDO/ DANIEL ENRIQUE VILLA ARIAS

DICIEMBRE DE 2020

UTP | Pereira

1 CONTENIDO

2 PRESENTACIÓN

La presente monografía presenta una introducción sobre los números de punto flotante y algunos ejemplos de cómo se emplean.

En los siguientes párrafos se presenta una descripción básica de que es y cómo se utilizan los números de punto flotante.

AUTORES: Gian Franco Posso Giraldo/ Daniel Enrique Villa Arias

CÓDIGO: 100468162/ 1010003883

CORREO: f.posso@utp.edu.co/ daniel.villla2@upt.edu.co

GITHUB:

3 NÚMEROS DE PUNTO FLOTANTE

Los números reales en base diez o decimal que nosotros habitualmente utilizamos para realizar cálculos matemáticos ya sea para tareas cotidianas o científicas generalmente son expresados en una de dos formas:

Una es en **punto fijo** (por ejemplo: $345,70 / 230Kg / -10°C, etc.), donde se emplean tres campos para la representación: **signo**, **parte entera** y **parte decimal**.

Otra forma es en **notación científica** o **punto flotante**, donde un número se expresa también con tres campos: **signo**, **mantisa** y **exponente** (por ejemplo: 10E-09 seg. = 0,000000001 seg. , 12.74E04 metros = 127400 metros, etc.).

Generalmente esta última notación se emplea cuando el número a representar es muy grande o muy pequeño y llevaría muchos dígitos su representación en punto fijo, lo cual puede resultar en errores de cálculo, de representación, etc.

Parte entera	Mantisa	Exponente	Notación científica	Valor en punto fijo
1	0.5	4	1.5×10^4	15000
5	0	-3	5×10^{-3}	0.005
6	0.667	-11	6.667E-11	0.0000000000667
-3	0.999	2	-3.999×10^3	-3999

4 SUMA Y RESTA

Cuando sumamos o restamos dos números en coma flotante se deben comparar los exponentes y hacerlos iguales, para lo cual hay que desplazar o alinear uno de ellos respecto al otro. Dados dos números en representación en coma flotante

$$x = m_x\, 2^{xe} \qquad y = m_y\, 2^{ye}$$

como:

las operaciones de suma y resta se definen de la siguiente forma, suponiendo que $X_e < Y_e$:

$x + y =$	$(m_x\, 2^{xe-ye} + m_y)\, 2^{ye} =$	$(m_x + m_y\, 2^{ye-xe})\, 2^{xe}$
$x - y =$	$(m_x\, 2^{xe-ye} - m_y)\, 2^{ye} =$	$(m_x - m_y\, 2^{ye-xe})\, 2^{xe}$

Ejemplo: Sean x e y los siguientes números enteros

x = 2560 y = 516000

en notación exponencial "normalizada" tendríamos

$x = 2.56\ 10^3$ $y = 5{,}16\ 10^5$

Si quisiéramos realizar las operaciones de suma o de resta en la notación en coma fija haríamos lo siguiente:

x+y ==>
```
   2560
+ 516000
  518560
```
que seria $5{,}1856\ 10^5$

x-y ==> (2560 - 516000) = -513440 que seria $5{,}13440\ 10^5$

Y para realizar estas operaciones en representación en coma flotante deberíamos aplicar las fórmulas vistas con anterioridad:

x+y ==> $x+y = 2.56\ 10^3 + 5{,}16\ 10^5$

$(2.56\ 10^{3-5} + 5{,}16)\ 10^5$

$(2{,}56\ 10^{-2} + 5{,}16)\ 10^5$

$(0{,}0256 + 5{,}16)\ 10^5$

$(5{,}1825)\ 10^5$

```
x-y  ==>  x-y =  2.56 10^3      -    5,16 10^5
                 (2.56 10^3-5   -    5,16)     10^5
                 (2,56 10^-2    -    5,16)     10^5
                 (0,0256        -    5,16)     10^5
                 (-5,1344)      10^5
```

5 DESBORDAMIENTO

Las operaciones de suma y resta, así como la multiplicación y la división pueden producir reboses, por producir resultados demasiado grandes (desbordamientos) o demasiado pequeños (subdesbordamientos). Hay cuatro tipos de reposes posibles:

-Desbordamiento del exponente. Es cuando un exponente positivo **E** excede de su valor máximo permitido. En algunos ordenadores el número **X** se representa entonces como +∞ ó -∞

-Subdesbordamiento del exponente. Es cuando un exponente negativo **E** excede de su valor máximo permitido. Esto significa que el número **X** es demasiado pequeño y se puede considerar como igual a **0**.

-Subdesbordamiento de mantisa. En el proceso de alineación de las mantisas, si los dígitos se desplazan hacia la derecha más allá de su bit menos significativo, lo que sucede es que se pierden y es como redondear el resultado.

-Desbordamiento de mantisa. En la suma de dos mantisas del mismo signo se puede producir un arrastre del bit más significativo. Esto se soluciona mediante la renormalización, desplazando a la derecha un bit la mantisa y ajustando el exponente.

6 RANGOS NÚMERICOS

Los números se almacenan en las variables. Una variable representa un trozo de la memoria del computador. La memoria está formada por una gran cantidad de bytes y cada byte está constituido por 8 bits. Un bit puede almacenar un 1 o un 0.

Se muestra una tabla con tipos de datos, tamaño que ocupan al ser almacenados y el intervalo de representación. Los mismos corresponden al lenguaje de programación Visual Basic 5.0

Tipo de datos	Tamaño de almacenamiento	Intervalo
Byte	1 byte	0 a 255
Boolean	2 bytes	**True** o **False**
Integer	2 bytes	-32.768 a 32.767
Long (entero largo)	4 bytes	-2.147.483.648 a 2.147.483.647
Single (coma flotante/ precisión simple)	4 bytes	-3,402823E38 a -1,401298E-45 para valores negativos; 1,401298E-45 a 3,402823E38 para valores positivos
Double (coma flotante/ precisión doble)	8 bytes	-1,79769313486232E308 a -4,94065645841247E-324 para valores negativos; 4,94065645841247E-324 a 1,79769313486232E308 para valores positivos
Currency (entero a escala)	8 bytes	-922.337.203.685.477,5808 a 922.337.203.685.477,5807
Decimal	14 bytes	+/-79.228.162.514.264.337.593.543.950.335 sin punto decimal; +/-7,9228162514264337593543950335 con 28 posiciones a la derecha del signo decimal; el número más pequeño distinto de cero es +/-0,0000000000000000000000000001

7 FORMATO IEEE 754

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

El estándar define:

- **Formatos aritméticos:** conjuntos de datos de punto 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 punto 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.

Este documento se enfocará en dos formatos, simple precisión (32 bits) y precisión doble (64 bits).

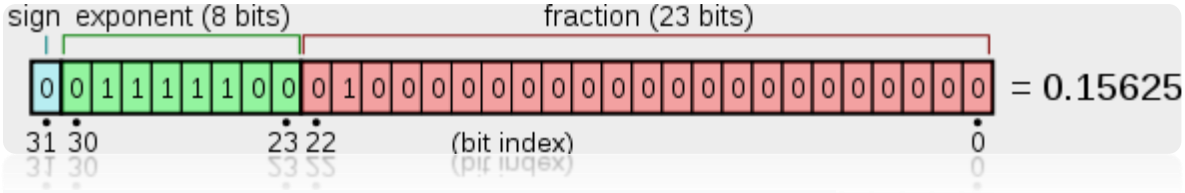
Simple precisión

El formato en punto flotante de simple precisión es un formato de número de computador u ordenador que ocupa 4 bytes (32 bits) en su memoria y representa un amplio rango dinámico de valores mediante el uso de punto flotante. En la norma o estándar IEEE 754 el formato de 32 bits de base 2 se conoce oficialmente como **binary32.**

El estándar IEEE 754 especifica que un formato **binary32** consta de:4

- **Bits de signo (S):** 1 bit.
- **Exponente desplazado (E):** 7 bits.
- **Significando o Mantisa (T):** 24 bits (23 almacenados explícitamente).

Este formato proporciona una precisión de 6 a 9 dígitos decimales significativos. Si una cadena decimal de hasta 6 dígitos decimales significativos se convierte en formato IEEE 754 de precisión simple y luego se convierte de nuevo al mismo número de dígitos decimales significativos, la cadena final debe coincidir con el original y si un número de precisión simple IEEE 754 se convierte en una cadena decimal con al menos 9 dígitos decimales significativos y luego se convierte de nuevo a un número de precisión simple, entonces el número final debe coincidir con el original



Estructura de un número en formato de coma flotante de simple precisión.

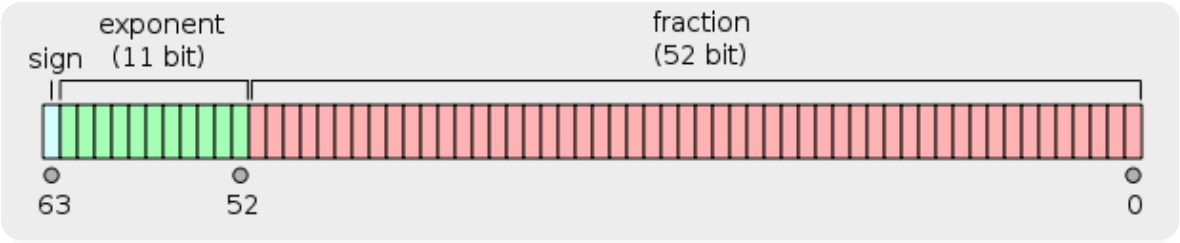
Doble precisión

El formato en coma flotante de doble precisión es un formato de número de computador u ordenador que ocupa 64 bits en su memoria y representa un amplio y dinámico rango de valores mediante el uso de la coma flotante. Este formato suele ser conocido como **binary64** tal como se especifica en el estándar IEEE 754.

El formato binario en coma flotante de doble precisión es de uso común en los computadores personales, debido a su rango más amplio de trabajo con respecto al formato en coma flotante de simple precisión, a pesar de su rendimiento y su coste de ancho de banda. Al igual que con el formato de coma flotante de simple precisión, carece de precisión en números enteros cuando se compara con un formato entero del mismo tamaño. El estándar IEEE 754 establece que un número en formato **binary64** consta de:

- **Bit de signo (S):** 1 bit.
- **Exponente (E**): 11 bits.
- **Significando o Mantisa**: 53 bits (52 bits son almacenados explícitamente).

Si una cadena decimal con un máximo de 15 dígitos significativos se convierte en una representación de doble precisión y luego se convierte de nuevo a una cadena con el mismo número de dígitos significativos, a continuación, la cadena final debe coincidir con la original. Si un número en doble precisión se convierte en una cadena decimal con al menos 17 dígitos significativos y luego se convierte de nuevo en doble, el número final debe coincidir con el original



Estructura de un número en formato de coma flotante de doble precisión.

8 CONCLUSIONES

Como la memoria de los ordenadores es limitada, no se puede almacenar números con precisión infinita, no importa si se usan fracciones binarias o decimales: en algún momento se tiene que parar, debido a esto la representación en punto flotante resulta muy útil para realizar estos cálculos que utilizan números o muy grandes o muy pequeños, demostrando ser una herramienta de gran valor a la hora de ser empleada, ya que facilita procedimientos y agiliza cálculos.

9 BIBLIOGRAFÍA

https://repl.it