Matemáticas III

Punto Flotante

Semana 06

Hermes Pantoja Carhuavilca
(hpantoja@utec.edu.pe)
Brigida Molina Carabaño
(bmolina@utec.edu.pe)
Rosulo Perez Cupe
(rperezc@utec.edu.pe)
Asistente: Victor Anhuaman
(vanhuaman@utec.edu.pe)

UTEC
UNIVERSIDAD DE INGENIERÍA & TECNOLOGÍA

Temas

1 PUNTO FLOTANTE

Logros de Aprendizaje

- Conocer las reglas de transformación de las representaciones de los números entre los sistemas de base 10 y base 2.
- Conocer las reglas internas, según las cuales las máquinas representan a los números del trabajo cotidiano.
- Representar los números en coma flotante.

Problema

Para resolver problemas numéricos con el computador, debemos trabajar con una cantidad de espacio finita y una precisión limitada.

- Los sistemas de precisión arbitraria todavía son costosos desde el punto de vista del tiempo de cómputo y desde el punto de vista del espacio requerido para su funcionamiento.
- Muchas veces hay una incertidumbre que no se puede eliminar en los valores de interés.
- Los modelos que se utilizan tienen una incertidumbre intrínseca por lo que no vale la pena mejorar la precisión más allá de cierto punto.

Es un hecho que debemos acostumbrarnos a vivir con los errores de representación, por lo que se hace necesario entender como ocurren.

Coma flotante

La representación de coma flotante es simplemente un tipo de notación científica que reconoce las limitaciones de espacio inherentes al computador.
El objetivo de la representación de coma flotante es ofrecer un sistema que sirva para las necesidades de cómputo modernas sin requerir que las operaciones se efectuen con precisión infinita.

IEEE-754

Los científicos que desarrollaron el cohete Ariane 5 vuelo 501 reutilizaron parte del código de su predecesor, el Ariane 4, pero los motores del cohete nuevo incorporaron también, sin que nadie se diera cuenta, un ”bug” en una rutina aritmética en la computadora de vuelo. Esto provocó, el 4 de junio de 1996, que la computadora fallara segundos después del despegue del cohete; 0,5 segundos más tarde falló el ordenador principal de la misión. El Ariane 5 se desintegró 40 segundos después del lanzamiento.

IEEE-754

Los errores en un sistema de coma flotante (*floating point* en inglés) pueden ser catastróficos:

- Vuelo inaugural del Ariane 5, 1996. (Video 4'26", Reporte)
- Batería anti-misiles *Patriot*, 1991. (Nota, Artículo)
- Error de division del Pentium, 1994. (Wikipedia)
- Elecciones parlamentarias en Alemania, 1992. (Nota)

El estándar IEEE-754-2008 reglamenta una cantidad de aspectos relevantes. Nosotros vamos a concentrarnos solo en algunos aspectos.

Actividad 1

P1

- Exprese el número 105 en el sistema de base 2
- Dado 11110101_2, expréselo en el sistema de base 10
- Exprese el número $0.8333\cdots$ en el sistema de base 2
- Cuál es el equivalente en base 10 para el número $101.11011011011\cdots_2$?

Solución:

Aritmética de Punto Flotante

Un sistema de punto flotante se especifica por la base β, el largo de la mantisa p, y límites para los exponentes de L y U.
El número x en punto flotante, en una base β tienen la forma:

$$x = \pm (1, d_1 d_2 \dots d_p)_\beta \beta^E$$

donde $d_1 d_2 \dots d_p$ es la mantisa, el 1er bit de la mantisa es implícito, y E es el exponente entero de la punto flotante.
Se define el sistema de punto flotante F con cuatro elementos

$$F(\beta, p, L, U) = F(base, precision, expomin, expomax)$$

$$L \leq E \leq U$$

$x \in \mathbb{R}$, $fl(x)$ es su punto flotante

Ejemplo

Dado el sistema de punto flotante $F(2, 4, -6, 7)$,

- Hallar la forma del sistema de punto flotante.
- Hallar el número de representaciones.
- El más pequeño en positivos.
- El más grande.
- ¿$0.5 \in F$?, ¿$\frac{3}{4} \in F$?, ¿$0.5 + \frac{3}{4} \in F$?

Cardinalidad

Cardinalidad de $F(\beta, p, L, U)$:

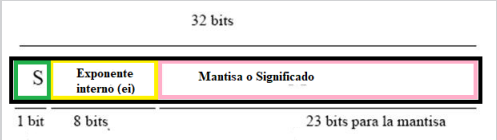
$$2(\beta - 1)\beta^{p-1}(U - L + 1) + 1$$

Example

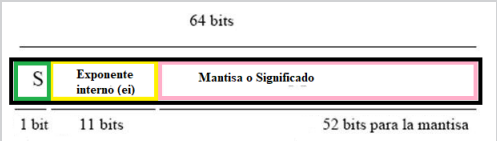
¿Cuántos números tendrá el sistema $F(2, 3, -1, 2)$?

Estándar IEEE-754

Simple Precisión: 32 bits

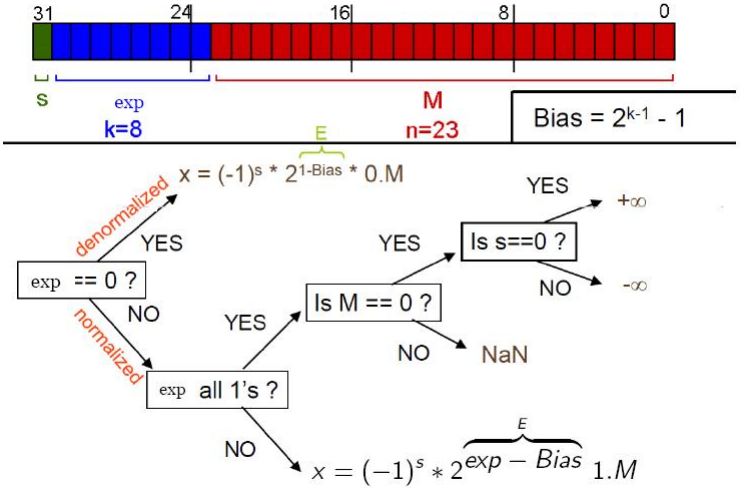


Doble Precisión: 64 bits



IEEE-754

La representación de punto flotante de Simple Precisión



Rango de Simple Precisión

- Los exponentes 00000000 y 11111111 son reservados.
- El valor más pequeño
  - Exponente exp: $00000001_2 = 1$
  - Exponente E: $E = exp - Bias = 1 - 127 = -126$
  - Parte fraccionaria: $000\ldots00$

$$x_{min} = 1.0 \times 2^{-126}$$

- El valor más grande
  - Exponente exp: $11111110_2 = 254$
  - Exponente E: $E = exp - Bias = 254 - 127 = +127$
  - Parte fraccionaria: $111\ldots11$

$$x_{min} = 1.111\ldots11 \times 2^{127}$$

Ejemplos

Ejemplo 1

Representa -3.75 en representación de simple precisión.

Solución:
Normalizamos de acuerdo a las normas de la IEEE 754.

- $-3.75 = -11.11_2 = -1.111 \times 2^1$
- *Bias* $= 127$, por lo que $127 + 1 = 128$ (es el exponente actual).

$$128 = 10000000_2$$

- El primer 1 en la mantisa es implícito, por lo que tenemos:

  1 10000000 11100000000000000000000

- Desde que tenemos implícito 1 en el significando, esto es igual a :
  $-1.111_2 \times 2^{(128-127)} = -1.111_2 \times 2^1 = -11.11_2 = -3.75$

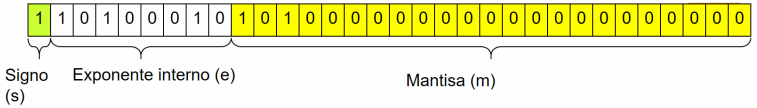
Continuación...

Ejemplo 2

Identifique el número de coma flotante correspondiente a la siguiente cadena de bits, utilizando precisión simple

1 10100010 10100000000000000000000

Solución:



$$Valor = (-1)^s \times (1, m)_2 \times 2^{e-127} = (-1)^1 \times (1, 10100000)_2 \times 2^{(10100010)_2 - 127}$$
$$= (-1) \times (1.625) \times 2^{162-127}$$
$$= (-1) \times (1.625) \times 2^{35} = -5.5834 \times 10^{10}$$

Actividad 2

P1

1. Represente el número -11.25 en el sistema IEEE-754 de punto flotante precisión simple
2. La secuencia de bits 11000010101100100000000000000000 representa a un número del sistema de base 10 según las reglas del sistema IEEE-754 de precisión simple. Halle dicho número.

Solución:

¿Qué es el Epsilón de la máquina?

El Epsilón de la máquina, es definido como el número más pequeño ϵ tal que $1 + \epsilon > 1$.
Es el número que mide la precisión de la máquina y por lo tanto el responsable del error de redondeo (numero de cifras decimales exactas). El número de máquina ϵ se define como la diferencia entre 1 y el sucesor de 1.

Actividad

Sea un sistema F basado en la norma IEEE-754 con las siguientes características:
Almacenamiento de 8 bits:

Signo	Exponente	Mantisa
1	3	4

Determine

- Dados los siguientes números: $a = 1.1101 \times 2^{-1} \in F$ y $b = 1.0101 \times 2^{2} \in F$. $a + b \in F$?
- El número 1
- El número $1 + \epsilon$
- El núméro ϵ
- El número 12.5
- El menor número positivo normalizado valor binario y decimal.
- El mayor valor positivo nrmalizado binario.
- El número Infinito $(-\infty)$ valor binario.
- Un NaN

IEEE-754

Algunos videos potencialmente interesantes sobre el formato IEEE-754:

- Una explicación de qué son y cómo funcionan y las razones por las que no son exactos por Computerphile.
- Un ejemplo de convertir un número decimal al estándar IEEE-754 a mano por Abishalini Sivaraman.
- Un ejemplo de convertir un número en el estándar IEEE-754 a un número decimal por Abishalini Sivaraman.

Gracias por su atención