Aritmética Computacional

Representação em Ponto Flutuante

- No mundo binário também existe limitações para representar números muito grande ou muito pequenos (fracionários).
- Para contornar isso os números decimais usam a notação científica
 - 976.000.000.000.000 = 9.76×10^{14}
 - $0,00000000000000976 = 9,76 \times 10^{-14}$
- Essa mesma técnica pode ser usada com números binários.

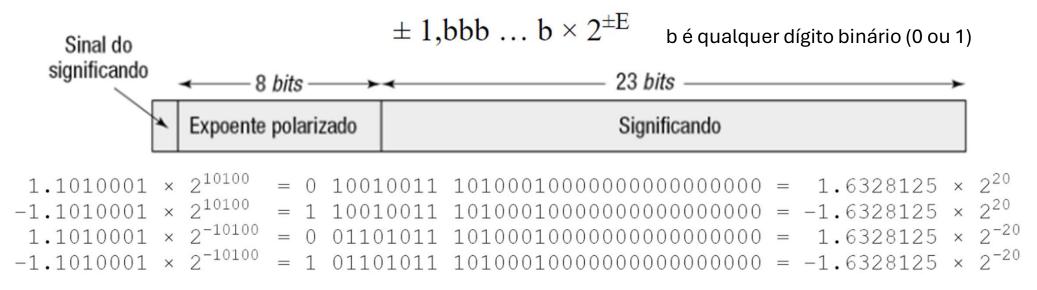
$$\pm S \times B^{\pm E}$$

- Esse número pode ser armazenado em uma palavra binária com três campos:
 - Sinal: mais ou menos
 - Significando S
 - Expoente E

- Exemplo de Formato Típico de 32 bits
 - Sinal
 - (0=positivo, 1=negativo)
 - Expoente
 - Representação Polarizada
 - Valor fixo (polarização) é subtraído desse campo para obter o verdadeiro valor do expoente
 - A polarização é dada por $(2^{k-1}-1)$, onde k é o número de bits do expoente
 - Para k=8 representa números de 0 a 255
 - Polarização é 127 ($2^7 1$), valores verdadeiros de -127 a +128;
 - Significando
 - Os 23 bits restantes



Exemplos Normalizados



- o sinal é armazenado no primeiro bit da palavra;
- o primeiro bit do verdadeiro significando é sempre 1 e não precisa ser armazenado no campo de significando;
- o valor 127 é acrescentado ao verdadeiro expoente para ser armazenado no campo de expoente;
- a base é 2.

• Limites de representação para uma palavra de 32 bits

