Ponto Flutuante e IEEE 754

Yuri Kaszubowski Lopes Éverlin Fighera Costa Marques

UDESC

Números reais e o Hardware

- Converter 0, 1₁₀ para base 2, qual o problema?
 - ▶ parte fracionária é uma repetição periódica
- Observe o programa abaixo:

```
1 #include<stdio.h>
2 #include<stdlib.h>
3
4 int main() {
5  float x = 0.0;
        for (int i = 0; i < 10; i++) {
    x += 0.1;
}</pre>
         printf("%.20f\n", x);
         return 0;
15 }
```

Saída:

1.00000011920928955078

Anotações

Anotações

Números reais e o Hardware

- É impossível representar qualquer número real na máquina
 - ► Temos uma infinidade de números reais, e o hardware é finito

 - Sempre há um número maior
 qualquer dois números reais há infinitos números reais
 - Armazenamos assim aproximações

Anotações		

YKL e EFCM (UDESC)

P0	nto	1-11	ПТП	ıar	1te

- Uma forma de armazenar essas aproximações é através de pontos flutuantes
- Conceito implementado em grande parte dos processadores comerciais
- Similar a notação científica normalizada
 - O número tem um e somente um digito antes da casa decimal, e não possui zeros antes da casa decimal (a IEEE 754 possui uma faixa de valores não normalizados também)
 - * $8.0_{10} \times 10^{-9}$ é normalizado * $0.8_{10} \times 10^{-8}$ **não** é normalizado * $80.0_{10} \times 10^{-10}$ **não** é normalizado
- Podemos fazer o mesmo com números binários
 - ▶ 1.0₂ × 2⁻¹ está normalizado
- Vamos chamar o ponto decimal de "ponto binário" para a base 2
- \bullet Exemplo: normalizar 1111, 112 = 1111, 112 \times 20
 - ► 111, 111₂ × 2¹

 - 11, 1111₂ × 2²

 ▶ 1, 11111₂ × 2³ ← Normalizado
 - Vamos exibir a base e a potência na base 10 na disciplina para simplificar a visualização

VV	1	ECI	м /	IDESC

Exercício Resolvido

Normalize os seguintes valores binários

- 11,11₂
- $\star = 1,111_2 \times 2^1$
- 3 111₂
- $\star\ = 1,11_2\times 2^2$ **3** 0,000001₂
 - $\star~=1,0_2\times 2^{-6}$

Anotações

Anotações

Anotações

Ponto Flutuante

- \bullet Para armazenar um valor binário normalizado arbitrário na memória, como 1,11111 $_2 \times 2^3$, quais campos são importantes?
 - Não precisamos armazenar o 1 antes do ponto binário, nem a base.

 Devemos armazenar:

 * Sinal (0 positivo, 1 negativo)

 * Expoente

 * mantissa (parte à direita do ponto binário)
- Os valores então sempre tem o formato:
 - ▶ ±1, mmmmmmm × 2^e
 - * mmmmmmm é a mantissa (ou fração)

 - * eeeeee é o expoente
 * Armazenamos apenas a mantissa, o expoente, e o sinal

Padrão IEEE 754 Anotações • Utilizado na maioria dos processadores comerciais x86-64, smartphones Define tamanhos para os campos de expoente e mantissa Precisão simples (float)Precisão dupla (double) Padrão IEEE 754 Anotações Precisão simples • Declarado como float em C 8 bits 23 bits Precisão dupla • Declarado como double em C (+) Consegue armazenar uma extensão maior de valores (+) Maior precisão devida a mantingo constituidad de valores • Quais as vantagens e desvantagens do double em relação ao float? (-) Ocupa mais memória (-) Pode precisar de mais ciclos do processador para efetuar cálculos

Overf	low e	Und	lerf	low
OVCII		Olic		

- Overflow: o expoente é muito grande para caber na memória
- Underflow: o expoente é muito pequeno para caber na memória

Anotações			

YKL e EFCM (UDESC) Ponto Flutuante

Valores Especiais

- Como podemos representar o número 0? Qual a dificuldade?
 Concordamos que o 1 antes do ponto binário é implícito, e não é representado pelo hardware
 ±1, mmmmmm × 2^{eeeeee}

Anotações

- Mas o número 0 em especial não tem 1 antes do ponto binário!
 Essas e outras exceções são tratadas com valores especiais.

Single p	recision	Double p	precision	Object represented			
Exponent	Fraction	Exponent	Fraction				
0 0		0	0	0			
0	Nonzero	0	Nonzero	± denormalized number			
1-254	Anything	1-2046	Anything	± floating-point number			
255	0	2047	0	± infinity			
255	Nonzero	2047	Nonzero	NaN (Not a Number)			

255	0	Nonzero	U	Nonzero	± denormalized number	
Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especialis definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especiais definidos no IEEE 754. Figura de Patierson e Hennessy (2017) Valores especialis de Number o IIII	1-254	Anything	1-2046	Anything	± floating-point number	
Al do Expoente E o sinal do expoente? Como representar? Exemplo: -1,11111 x 2-2 -1.11111 x 2-2 -1.1111 x 2-2 -1.1111 x 2-						
Al do Expoente E o sinal do expoente? Como representar? Exemplo: -1, 11111 × 2-2	255					
al do Expoente E o sinal do expoente? Como representar? Exemplo: -1, 111111 × 2-2 -1,111111 × 2-2 -1,111111 × 2-2 31, 30, 29, 28, 27, 29, 25, 24, 23, 22, 21, 20, 39, 18, 17, 16, 16, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 10, 10, 10, 10, 10, 10, 10, 10		Valores especiais	definidos no IEE	E 754. Figura de Pa	itterson e Hennessy (2017)	
al do Expoente E o sinal do expoente? Como representar? Exemplo: -1, 111111 × 2-2 -1,111111 × 2-2 -1,111111 × 2-2 31, 30, 29, 28, 27, 29, 25, 24, 23, 22, 21, 20, 39, 18, 17, 16, 16, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 10, 11, 11, 11, 11, 12, 11, 10, 10, 10, 10, 10, 10, 10, 10, 10				-		
E o sinal do expoente? Como representar? Exemplo: -1,11111 x 2 ⁻² -1,11111 ₂ x 2 ² -1,11111 ₂ x 2 ² 33 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4	YKL e EFCM (U	DESC)	Ponto	Flutuante		10/19
31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent 1 bit 8 bits 23 bits -1.111112 × 2² 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent 1 bit 8 bits 23 bits -1.111112 × 2² 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent 1 bit 8 bits 23 bits -1.111112 × 2² 23 bits -1.111112 × 2² 24 20 20 21 20 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 Fonto Flutusinte Ponto Flutusinte Ponto Flutusinte Ponto Flutusinte Ponto Flutusinte 11/15 E 754: notação com bias Biased Expoent E 754: notação com bias Dias é o valor intermediário entre todos os possíveis de serem representados no expoente 12710 (0111 11112) para precisão simples 102310 (011 1111 11112) para precisão dupla O expoente é somado ao bias E Expoente —1 10 na notação com bias se torna:	E o sinal	do expoente	? Como rep : 2 ⁻²	oresentar?		
s exponent 1 bit 8 bits 23 bits -1:111112 × 2² 31 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent 1 bit 8 bits 23 bits -1:111112 × 2² 31 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent -1:11112 × 2² 31 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent -1:11112 × 2² -1:11112 × 2² -1:11112 × 2² -1:111112 × 2² -1:1						
1 bit 8 bits 23 bits -1.111112 × 2² 31, 30 29 28 27 28 25 24 23 22 21 20 19 18 37 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent 1 bit 8 bits 23 bits -1.111112 × 2² 31, 30 29 28 27 28 25 24 23 22 21 20 19 18 37 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 s exponent VKL s EFCM (UDESC) Ponto Flutusints 11/15 Ponto Flutusints			2 21 20 19 18			
s exponent 1 bit 8 bits 23 bits -1.111112 × 2² 31, 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 -1.111112 × 2² 31, 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 Fraction 11/15 Ponto Flutuente 11/15 Fraction 11/15 Ponto Flutuente 11/15 Ponto Flutuente 11/15 Ponto Flutuente Ponto Flutuente Ponto Flutuente Ponto Flutuente 11/15 Ponto Flutuente	1 bit					
s exponent 1 bit 8 bits 23 bits -1.111112 × 2² 31, 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 -1.111112 × 2² 31, 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 Fraction 11/15 Ponto Flutuente 11/15 Fraction 11/15 Ponto Flutuente 11/15 Ponto Flutuente 11/15 Ponto Flutuente Ponto Flutuente Ponto Flutuente Ponto Flutuente 11/15 Ponto Flutuente		-1 11111 × 2·2				
s exponent 1 bit 8 bits 23 bits -1:111112 × 2 ² 31 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 Ponto Flutuante Ponto Flutuante Ponto Flutuante Ponto Flutuante Ponto Flutuante 11/18 Ponto Flutuante Fraction 11/18 Ponto Flutuan) - <u></u>				
s exponent 1 bit 8 bits 23 bits -1:111112 × 2 ² 31 30 29 28 27 28 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 Ponto Flutuante Ponto Flutuante Ponto Flutuante Ponto Flutuante Ponto Flutuante 11/18 Ponto Flutuante Fraction 11/18 Ponto Flutuan						
A solution of the component of the compo			2 21 20 19 18			
Ou então sinal e magnitude, como na mantissa mas não! O IEEE 754 especifica que o expoente utiliza uma notação com bias Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 12710 (0111 11112) para precisão simples 1270 (0111 11112) para precisão dupla O expoente é somado ao bias Expoente —110 na notação com bias se torna:						
Ponto Futuante				231	its	
Poderíamos utilizar complemento de 2 Ou então sinal e magnitude, como na mantissa mas não! O IEEE 754 especifica que o expoente utiliza uma notação com bias Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 12710 (0111 11112) para precisão simples 102310 (011 1111 11112) para precisão dupla O expoente é somado ao bias Expoente — 110 na notação com bias se torna:		-)1,11111 ₂ × 2 ²				
Poderíamos utilizar complemento de 2 Ou então sinal e magnitude, como na mantissa mas não! O IEEE 754 especifica que o expoente utiliza uma notação com bias Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 12710 (0111 11112) para precisão simples 102310 (011 1111 11112) para precisão dupla O expoente é somado ao bias Expoente — 110 na notação com bias se torna:						
Ponto Flutuante Poderíamos utilizar complemento de 2 Ou então sinal e magnitude, como na mantissa mas não! O IEEE 754 especifica que o expoente utiliza uma notação com bias • Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente • 127₁₀ (0111 1111₂) para precisão simples • 1023₁₀ (011 1111 111₂) para precisão dupla O expoente é somado ao bias • Expoente —1₁₀ na notação com bias se torna:	31 30 29 28	27 26 25 24 23 2	2 21 20 19 18	17 16 15 14 13 1	2 11 10 9 8 7 6 5 4 3 2 1 0	
Poderíamos utilizar complemento de 2 Ou então sinal e magnitude, como na mantissa mas não! O IEEE 754 especifica que o expoente utiliza uma notação com bias • Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente • 127 ₁₀ (0111 1111 ₂) para precisão simples • 102 ₁₀ (0111 1111 ₂) para precisão dupla O expoente é somado ao bias • Expoente —1 ₁₀ na notação com bias se torna:				fract	ion	
Ou então sinal e magnitude, como na mantissa mas não! O IEEE 754 especifica que o expoente utiliza uma notação com bias • Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente • 127 ₁₀ (0111 1111 ₂) para precisão simples • 1023 ₁₀ (0111 1111 1111 ₂) para precisão dupla O expoente é somado ao bias • Expoente —1 ₁₀ na notação com bias se torna:	YKL e EFCM (U	DESC)	Ponto	Flutuante		11/19
mas não! O IEEE 754 especifica que o expoente utiliza uma notação com bias ► Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente ► 127₁₀ (0111 1111₂) para precisão simples ► 1023₁₀ (011 1111 111₂) para precisão dupla O expoente é somado ao bias ► Expoente −1₁₀ na notação com bias se torna:	Poderíam	os utilizar co	mplemento	de 2		
O IEEE 754 especifica que o expoente utiliza uma notação com bias • Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente • 127 ₁₀ (0111 1111 ₂) para precisão simples • 102 ₁₀ (0111 1111 ₂) para precisão dupla O expoente é somado ao bias • Expoente -1 ₁₀ na notação com bias se torna:			nitude, com	o na mantis:	sa	
 ▶ Biased Expoent E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente ▶ 127₁₀ (0111 1111₂) para precisão simples ▶ 1023₁₀ (011 1111 1111₂) para precisão dupla O expoente é somado ao bias ▶ Expoente -1₁₀ na notação com bias se torna: 				oonto utili-o	uma nataoão com bica	
E 754: notação com bias O bias é o valor intermediário entre todos os possíveis de serem representados no expoente ► 127₁₀ (0111 1111₂) para precisão simples ► 1023₁₀ (011 1111 1111₂) para precisão dupla O expoente é somado ao bias ► Expoente −1₁₀ na notação com bias se torna:			a que o exp	oente utiliza	uma notação com bias	
O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 127 ₁₀ (0111 1111 ₂) para precisão simples 1023 ₁₀ (011 1111 1111 ₂) para precisão dupla O expoente é somado ao bias Expoente —1 ₁₀ na notação com bias se torna:		Pro-				
O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 127 ₁₀ (0111 1111 ₂) para precisão simples 1023 ₁₀ (011 1111 1111 ₂) para precisão dupla O expoente é somado ao bias Expoente —1 ₁₀ na notação com bias se torna:						
O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 127 ₁₀ (0111 1111 ₂) para precisão simples 1023 ₁₀ (011 1111 1111 ₂) para precisão dupla O expoente é somado ao bias Expoente —1 ₁₀ na notação com bias se torna:						
O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 127 ₁₀ (0111 1111 ₂) para precisão simples 1023 ₁₀ (011 1111 1111 ₂) para precisão dupla O expoente é somado ao bias Expoente —1 ₁₀ na notação com bias se torna:						
O bias é o valor intermediário entre todos os possíveis de serem representados no expoente 127 ₁₀ (0111 1111 ₂) para precisão simples 1023 ₁₀ (011 1111 1111 ₂) para precisão dupla O expoente é somado ao bias Expoente —1 ₁₀ na notação com bias se torna:	EE 754:	: notação	com bi	as		
representados no expoente 127 ₁₀ (0111 1111 ₂) para precisão simples 1023 ₁₀ (011 1111 1111 ₂) para precisão dupla O expoente é somado ao bias Expoente -1 ₁₀ na notação com bias se torna:					nana(vala de	
 127₁₀ (0111 1111₂) para precisão simples 1023₁₀ (011 1111 1111₂) para precisão dupla O expoente é somado ao bias Expoente −1₁₀ na notação com bias se torna: 				ie lodos os	possiveis de serem	
 ▶ 1023₁₀ (011 1111 1111₂) para precisão dupla O expoente é somado ao bias ▶ Expoente −1₁₀ na notação com bias se torna: 	-	-		ão simples		
O expoente é somado ao bias Expoente -1 ₁₀ na notação com bias se torna:					l	
Expoente −1₁0 na notação com bias se torna:						
	► Expo	ente -1 ₁₀ na r	notação com			

- ★ -1₁₀ + 127₁₀ = 126₁₀ = 011111110₂ (precisão simples)
 ★ -1₁₀ + 1023₁₀ = 1022₁₀ = 01111111110₂ (precisão dupla)
 Expoente 1₁₀ na notação com bias se torna:
 ★ 1₁₀ + 127₁₀ = 128₁₀ = 1000 0000₂ (precisão simples)
 ★ 1₁₀ + 1023₁₀ = 1024₁₀ = 100 0000 0000₂ (precisão dupla)
- A notação com bias foi feita para tornar ordenações via hardware mais rápidas e simples

Anotações			
, ποιαγούδ			
A			
Anotações			
	·		

IEEE 754

• Um ponto flutuante é então representado por:

(-1)sinal × (1.0 + mantissa) × 2(expoente-bias) Somente esses valores são armazenados

- Exemplo: Representar −0,75₁₀ em precisão simples
 - $ightharpoonup -0.75_{10}$ pelo método das multiplicações nos dá $-0,11_2$

 - Normalizando, -0, 11₂ = -1, 1₂ × 2⁻¹
 A mantissa é então 1 (somente a parte fracionária)
 O expoente é -1₁₀ + 127₁₀ = 126₁₀ = 0111 1110₂
 - O bit de sinal é 1 (negativo)
 - Assim:

1	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1 bi	1 bit 8 bits 23 bits																														
	Figura de Patterson e Hennessy (2017)																														

- Como ficaria em precisão dupla?

 - mantissa com 52 bits (neste caso só zeros a direita)
 recalcular expoente: −1₁0 + 1023₁0 = 1022₁0 = 011 1111 1110₂

YKL	- 6	COL	// UD	ECO

Anotações

Anotações

Exemplo

• Converta o valor representado em precisão simples para decimal

- O expoente é 1000 0001₂ = 129₁₀

 - A mantissa 6 1,01 $_2$ = 1 × 2 0 + 0 × 2 $^{-1}$ + 1 × 2 $^{-2}$ = 1 $_{10}$ + 0,25 $_{10}$ = 1,25 $_{10}$ O bit de sinal 6 1 (negativo) Logo temos: -1,25 × 2 2 = -5₁₀

YKL	e EFCM	(UDESC)

Arredondamento

- A parte fracionária pode não caber na mantissa
 - O processador utilizará técnicas de arredondamento
 Perda de precisão
- O IEEE 754 define 4 formas de arredondamento, que podem ser setadas no processador:

 - Arredondar para cima (teto)Arredondar para baixo (piso)
 - Truncar (ignorar as demais casas)
 - Nearest even (par mais próximo) (mais utilizado)
- Requerem que o processador mantenha alguns bits extras para gerência
- Detalhes sobre o arredondamento e suas implementações, são encontrados em Patterson e Hennessy (2017) e Hennessy e Patterson (2014).

Anotações			

IEEE 754 Anotações • O padrão para precisão simples e dupla é embutido no hardware Independe de linguagem • Hardwares simples podem não implementar o padrão e.g., microcontroladores Nesse caso, o padrão é implementado via software Exercícios extras Anotações Qual o maior e o menor valor que podem ser representados em ponto flutuante de precisão dupla e simples (desconsiderando infinito)? Quais são seus equivalentes em decimal? 2 Exiba os seguintes valores em ponto flutuante. Quando necessário trunque (ignore os bits que não couberem) os valores. Faça o desenho da memória como nos exemplos e coloque os endereços dos bits (para deixar claro a ordem dos bits). ► -16,015625₁₀ Em precisão simplesEm precisão dupla $-0, 1_{10}$ Em precisão simples Em precisão dupla ▶ 0, 1₁₀ ★ Em precisão simples ★ Em precisão dupla ▶ 0,125₁₀ Em precisão simples Em precisão dupla Em meia precisão: 10 bits para mantissa, 5 para expoente e 1 para sinal Referências Anotações • TOCCI, R.J.; WIDMER, N.S. Sistemas digitais: princípios e aplicações. 11a ed, Prentice-Hall, 2011. • RUGGIERO, M.; LOPES, V. da R. Cálculo numérico: aspectos teóricos e computacionais. Makron Books do Brasil, 1996. • NULL, L.; LOBUR, J. Princípios Básicos de Arquitetura e Organização de Computadores. 2014. Bookman, 2009. ISBN 9788577807666.

YKL e EFCM (UDESC) Ponto Flutuante