

Pontifícia Universidade Católica do Rio Grande do Sul Instituto de Informática Organização de Computadores - GAPH

Unidade 3 Aritmética Computacional Uma breve introdução

Autores: Fernando Gehm Moraes & Ney Laert Vilar Calazans

Porto Alegre, junho de 1998

Última revisão: 30/novembro/2001





Introdução





Introdução a Aritmética Computacional

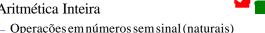
- Ø Bibliografia:
 - Patterson, D. A. "Computer Architecture: a quantitative approach", Morgan Kaufmann Pubs., 1996. Appendix A.
 - Texas Instruments. "TMS320C4x User's Guide", 1996. Capítulo 5. Disponível na Internet (formato PDF):
 - » http://www-s.ti.com/sc/psheets/spru063b/spru063b.pdf
 - » ftp://ftp.inf.pucrs.br/pub/calazans/texas/spru063b.pdf

Operações sobre naturais - Soma

- Ø Hardware mais simples baseado em ripple-carry (propagação em onda do vai-um) e componentes simples (meio somador e somador completo)
- - meio-somador: $s_i = a_i \oplus b_i$, $c_{i+1} = a_i \wedge b_i$;
 - somador completo: $s_i = a_i \oplus b_i \oplus c_i$ $c_{i+1} {=} a_i {\wedge} b_i {\vee} a_i {\wedge} c_i {\vee} b_i {\wedge} c_i.$

Sumário

- Ø Introdução
- Aritmética Inteira



- » soma, subtração, multiplicação e divisão
- Representações de números com sinal (inteiros)
- Aritmética não Inteira (racionais)
 - Representação de racionais
 - Operações com o padrão IEEE-754



Introdução a Aritmética Computacional

- Ø Uma parte especializada do projeto de computadores
- Ocontudo, uma parte muito, muito importante:
 - gráficos, comunicações, transações bancárias, matemática computacional, cálculo de estrutura, solução de equações, entre tantas aplicações;
- Pentium(otimizou errado um PLA usado em ÷);
- Aqui, revisão estendida de inteiros e padrão para números racionais (IEEE-754).

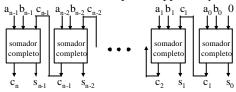
Sumário

- ✓ Introdução
- Aritmética Inteira
 - Operações em números sem sinal (naturais)
 - » soma, subtração, multiplicação e divisão



Operações sobre naturais - Soma

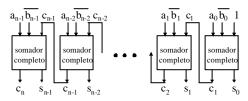
Ø Estruturado somador completo "ripple-carry":



- \bigcirc Problema: atraso de geração do vai-um = O(n), onde n é o número de bits do somador;
- \emptyset Hardware adicional -> atraso = $O(\log n)$.

📕 Operações sobre naturais - Subtração

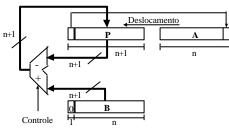
Ø Estruturado subtrator completo a-b "ripple-carry":



- Ø Mesmo problema de atraso, solução análoga;
- Ø Problema adicional, operação não definida se b>a.

📕 Operações sobre naturais - Divisão

Solução para a/b: subtrações sucessivas, n passos;



Divisão A/B - Exemplo

B = 00101 (5)

A cada passo, mostr 1) - primeira linha e 4) - segunda linha

2) PO P-D; 3) If (P<0), A₀=0 else A₀=1; 4) If (P<0), restaura Pfazendo PÜ P+B

1) desloca P&A p/ esq 1 bit;

		P (conterá o resto)						conte	erá a	divisa	io)
passo	0	0	0	0	0	0	1	1	0	1	1
1	0	0	0	0	0	1	1	0	1	1	0
	0	0	0	0	0	1	1	0	1	1	0
2	0	0	0	0	1	1	0	1	1	0	0
	0	0	0	0	1	1	0	1	1	0	0
3											
4					•						
5											

Divisão A/B - Exemplo

A = 11011 B = 00101(5)

A cada passo, mos 1) - primeira linha 4) - segunda linha

1) desloca P&A p/ esq 1 bit;

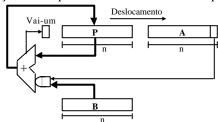
3) If (P<0), A₀=0 else A₀ =1; 4) If (P<0), restaura Pfazendo PÜ P+B

A (conterá a divisão)

1 1 0 1 1 1 1 0 0 1 1 0 0 1 0 1 1 0 1 1 0 0 1 1 0 0 0 1 1 0 1 4 0 0 1 0 0

Operações sobre naturais - Multiplicação

Ø Soluçãonaturalparaa*b: somas sucessivas n passos;



- Ø Inicialmente, P=0, A=a, B=b. Cada passo, duas partes:
 - soma carregada em P;
 - P & A deslocado um bit para a direita.

Divisão A/B - Exemplo

A = 11011 (27)

B = 00101(5) A cada passo, mostra 1) - primeira linha e 4) - segunda linha

1) desloca P&A p/ esq 1 bit; 2) PÜ P-B; 3) If (P<0), A₀=0 else A₀ =1; 4) If (P<0), restaura Pfazendo PÜ P+B

		P	(con	erá c	rest	o)	A (conterá a divisão)					
asso	0	0	0	0	0	0	1	1	0	1	1	
1	0	0	0	0	0	1	1	0	1	1	0	
	0	0	0	0	0	1	1	0	1	1	0	
2												
3												
4												
5												

Divisão A/B - Exemplo

B = 00101 (5)
A cada passo, mostro
1) - primeira linha e
4) - segunda linha

1) desloca P&A p/ esq 1 bit; 3) If (P<0), A₀=0 else A₀ =1; 4) If (P<0), restaura Pfazendo PÜ P+B

		Р	P (conterá o resto)					A (conterá a divisão)				
passo	0	0	0	0	0	0	1	1	0	1	1	
1	0	0	0	0	0	1	1	0	1	1	0	
	0	0	0	0	0	1	1	0	1	1	0	
2	0	0	0	0	1	1	0	1	1	0	0	
	0	0	0	0	1	1	0	1	1	0	0	
3	0	0	0	1	1	0 /	1	1	0	0	0	
	0	0	0	0	0	1/	1	1	0	0	1	
4												
5												
						I /	_					

Divisão A/B - Exemplo

A = 11011

B = 00101(5)

1) desloca P&A p/ esq 1 bit

3) If (P<0), A₀=0 else A₀ =1; 4) If (P<0), restaura Pfazendo PÜ P+B

		Р	(con	terá c	rest	o)	Α(conte	erá a	divis	io)
passo	0	0	0	0	0	0	1	1	0	1	1
1	0	0	0	0	0	1	1	0	1	1	0
	0	0	0	0	0	1	1	0	1	1	0
2	0	0	0	0	1	1	0	1	1	0	0
	0	0	0	0	1	1	0	1	1	0	0
3	0	0	0	1	1	0	1	1	0	0	0
	0	0	0	0	0	1	1	1	0	0	1
4	0	0	0	0	1	1	1	0	0	1	0
	0	0	0	0	1	1	1	0	0	1	0
5	0	0	0	1	1	1	0	0	1	0	0
	0	0	0	0	1	0	0	0	1	0	1

Resto = 2

Quociente = 5

Operações sobre naturais - Divisão

- Ø Algoritmo: versão binária procedimento lápis e papel;
- Ø Existe versão sem restauração de P (ver H & P);
- ∅ n passos, somador maior que na multiplicação;
- Ø Deve-se testar se divisor =0!
- Ø Restauração desnecessária se teste na saída do somador/subtrator, bem como somador.

10/hovembro/2001 - Moraes e Ney

Representações de Inteiros

- Quatro métodos comuns (convenções) p/ inteiros:
 - sinal-magnitude (SM) bit mais significativo é sinal, restante é o valor absoluto equivalente a binário sem sinal;
 - complemento de 1 (1's) positivos, idem a SM, negativos são positivos com valor invertido;
 - complemento de 2 (2's) positivos, idem a SM, negativos obtidos adicionando 1 a 1's;
 - polarização (bias) representação é a do positivo binário obtido a partir da soma de um valor k.

visio em 30hovembro0001 - Moraes e Ney

🔣 Transbordo em Inteiros (Overflow)

O Casos de transbordo em complemento de 2 (2's)

- Para 5 bits, faixa representável é -16 a +15

Decimal	Binário	Decimal	Binário	Decimal	Binário
Vai-um:	00111	Vai-u	m: <mark>11</mark> 011	Vai-u	ım:00001
+ 5	00101	- 5	11011	+ 5	00101
+ 7	00111	- 7	11001	<u>- 7</u>	11001
+ 7 + 12	01100	-12	10100	- 2	11110

tics: revisio em 30hovembro/2001 - Moraes e Ney

Sumário

- ✓ Introdução
- ✓ Aritmética Inteira
 - ✓ Operações em números sem sinal (naturais)
 - ✓ soma, subtração, multiplicação e divisão
 - ✓ Representações de números com sinal (inteiros)
- Aritmética não Inteira (racionais)
 - Representação de racionais



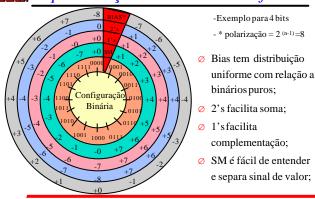
| | | | Sumário

- ✓ Introdução
- Aritmética Inteira
 - ✓ Operações em números sem sinal (naturais)
 - ✓ soma, subtração, multiplicação e divisão
 - Representações de números com sinal (inteiros)

Gaph

a: revisão em 30/novembro/2001 - Moraes e Ney

Representações de Inteiros - 4 formas



Transbordo em Inteiros (Overflow)

Ø Casos de transbordo em complemento de 2 (2's)

- Para 5 bits, faixa representável é -16 a +15

Decimal Vai-um:	Binário 00111	Decimal Vai-u	Binário m: <mark>11</mark> 011	Decima Vai	l Binário -um: <mark>00</mark> 001
+ 5 + 7 +12	00101 00111 01100	- 5 - 7 -12	11011 11001 10100	+ 5 - 7 - 2	00101 11001 11110
Decimal	Binário			ositivoene	
Vai-um - 5 + 7 + 2	:11111 11011 00111 00010	Decimal Vai-um + 8 + 9 +17	Binário n: 01000 01000 <u>01001</u> 10001	Decima Vai - 8 - 9 -17	1 Binário -um:10000 11000 <u>10111</u> 01111
Xor de 2 últin	nos vai-uns ider	ntifica transbord	o. Negativ	vo!	Positivo

Aritmética não Inteira (racionais)

- Muitasaplicações requerem números não-inteiros:
 - $-\ matem\'atica computacional, engenharia,\ etc;$
- Ø Racionais (Q):representados como fração a/b, a e b inteiros; Irracionais (I): têm mantissa infinita sem repetição (e=2.7218... eπ=3.14..., por exemplo);
- Ø Reais: Q ∪ I; I e R não representáveis em computadores, porquê?
- Ø Aproximação de reais em computadores: Q.

🌃 Aritmética não Inteira - representações

- Ø Primeiros computadores ponto (ou vírgula) fixo;
- Ø Hoje, apenas ponto flutuante; representações possuem um significandos (a mantissa do número), um expoentee e uma base b, e o número é dado por: sxb^e;
- Antes, muitos formatos, hoje, padrão quase universal: definido pela IEEE (IEEE-754-1985), quatro formatos, dois fixos (precisão simples, SP e precisão dupla, DP) e dois variáveis (precisão simples, SE, e dupla, DE, estendidas). Igual ao padrão internacional IEC-559.

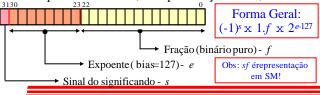
evisão em 30/novembro/2001 - Moraes e Ney

25

Uma Representação de racionais - IEEE-754-1985

- - 1 bit para sinal do significando s;
 - significando com 24 bits de precisão, primeiro sempre 1, exceto quando denormalizado, onde é 0 (1° bit implícito);

expoente e de 8 bits, com polarização = 127;



IEEE-754-1985 - Valores especiais e denormalização

- Principal novidade do padrão é existência de valores especiais, que permitem tratar exceções:
 - valores infinitos;
 - divisão por zero;
 - operações para os quais os reais não são fechados, como raiz quadrada;

revisão em 30/novembro/2001 - Moraes e Ney

20

IEEE-754-1985 - Parâmetros e outros formatos

Os diferentes formatos e valores para os parâmetros que os definem:

formato -> parâmetros	precisão simples (SP)	precisão simples estendida (SE)	precisão dupla (DP)	precisão dupla estendida (DE)
bits de precisão (p)	24	≥32	53	≥64
Emax	127	≥1023	1023	≥16383
Emín	-126	≤-1022	-1022	≤-16382
Polarização (bias)	127	depende	1023	depende
Total de bits	32 exatamente	variável, ≥43, <64	64 exatamente	variável, ≥79

Uma Representação de racionais - IEEE-754-1985

- Ø Diferenças para formatos anteriores à padronização:
 - ao arredondar um resultado no meio da faixa, pega valor par mais próximo;
 - inclui valores especiais:
 - » NaN Not a number (ex: raiz de negativo);
 - » -∞ e +∞ Mais ou menos infinito (ex: -1/0 e +1/0);
 - números denormalizados para resultados com valor menorque +1.0×2^{Emínimo};
 - arredonda para mais próximo (default) mas tem mais três modos de arredondamento;
 - recursos sofisticados paralidar com exceções.

IEEE-754-1985 - Exemplos

1) Número positivo, maior que zero

Forma Geral: $(-1)^s \times 1.f \times 2^{e-127}$

$$((-1)^0 \times 1.01 \times 2^{129 \cdot 127})_2 = (+1.01 \times 2^2)_2 = (+5)_{10}$$

2) Número negativo, menor que zero

 $((-1)^{I} \times 1.0111 \times 2^{126-127})_{2} = (-0.10111)_{2} = (-0.71875)_{10}$

Aritmética: revisão em 30/novembro/2001 - Moraes e Ney

28

IEEE-754-1985 - Valores especiais e denormalização

∅ Há 5 casos que definem o valor de número em algum formato do padrão IEEE-754 (exemplo para SP):

1) e=255, f≠0	v=NaN (not a number)
2) e=255, f=0	v=(-1) ^s ∞
3) 0 <e<255< td=""><td>$v=(-1)^s x (1.f) x 2^{e-127}$</td></e<255<>	$v=(-1)^s x (1.f) x 2^{e-127}$
4) e=0, f≠0	$v=(-1)^s x (0.f) 2^{-126}$
5) e=0, f=0	$v=(-1)^{s} x (0)$ (zero)

✓ Introdução



✓ Aritmética Inteira

Sumário

- ✓ Operações em números sem sinal (naturais)
 ✓ soma, subtração, multiplicação e divisão
- ✓ Representações de números com sinal (inteiros)
- Aritmética não Inteira (racionais)
 - ✓ Representações de racionais
 - Operações com o padrão IEEE-754



- Operação mais fácil de se implementar em hardware:
 - multiplicação e não soma
 - » adaptação de significandos devido a diferentes expoentes complica soma;
 - » multiplicação é direta, multiplica significandos e soma expoentes, exceto quando resultado é caso especial;
- Veremos aqui introdução a ambos, multiplicação e soma.

33

Operação de multiplicação com o padrão IEEE-754


```
1 10000010 000000000000000000000 = -1x2^3 = -8 0 10000011 000000000000000000000 = 1x2^4 = 16
```

- -1)Desempacotando- $> 1.0 \times 1.0 = 1.0$
- » logo, resultado tem a forma:

- 2) Expoente - fórmula para cálculo do expoente:

```
(exp polarizado(e<sub>1</sub>+e<sub>2</sub>))<sub>2's</sub> = (exp polarizado(e<sub>1</sub>) + (exp polarizado(e<sub>2</sub>) + (-polarização)<sub>2's</sub>, ou seja,

10000010 = 130
10000011 = 131
```

10000010 = 130 10000011 = 131 +10000001 = -127 10000110 = 134-0

10000110 = 134=e -> E=134-127=7!

35

Operação de multiplicação com o padrão IEEE-754

- 3) arredondamento (continuação)
 - » se *r* é menor que 5 resultado pronto;
 - » se r é maior que 5 soma-se 1 ao número em negrito;
 - » se arredondador exatamente 5 (em binário, 1) examinar bits seguintes, até achar um diferente de 0 ou chegar ao fim:
 - técnica-usa o "bit grudento" (sticky bit), durante a multiplicação, o
 OU lógico de todos os bits a partir do bit r;
 - Caso 1 desloca um bit p/ esq; Caso 2 incrementa expoente.

Produto(p=6)
Caso 1: x0=0
Caso 2: x0=1



27

IEEE-754-1985 - Multiplicação e denormalização

- Ocntrolar underflow é complicado, devido aos denormalizados:
 - porexemplo, 1x2-⁶⁴ X 1x2⁶⁵ é 1x2-¹²⁹, expoente não representável no formato normalizado, mas representável como 0.125x2-¹²⁶;
 - se o expoente não polarizado for menor que -126, produto deve ser deslocado bit a bit e expoente incrementado até -126. Se todo o significando se anular, aí sim, houve underflow;
- Quando um dos operandos é denormalizado, geram-se 0s à esquerda;
- Ø Denormalizados causam montes de problemas p/ multiplicação;
- Multiplicadores de alto desempenho ignoram denormalizados, gerando exceções para o software cuidar;
- Denormalizados não são freqüentes, perda em tempo é estatisticamentepequena.

Ø 3Passos:

- multiplicar significandos (não fração, desempacotaro número) usando multiplicação inteira, sem sinal (SM);
- calcular expoente
 - » lembrar da polarização.
- arredondamento devido ao aumento da precisão após operação;

Aritmética: revisão em 30/novembro/2001 - Moraes e Ney

3.4

Operação de multiplicação com o padrão IEEE-754

- 3) arredondamento precisão é importante:
- » casos de arredondamento (em decimal, análogo a binário):
- a) 1.23 r=9, 9>5, então arredonda p/8.34 \times 6.78 \times 8.3394
- c) 1.28 r=6, 6>5, então arredonda p/ 1.00×10^{1} $\frac{x \ 7.81}{0 \ 9.99 \ 68}$
 - » em binário, meio da faixa (5 em decimal) é dígito 1!
 - » negrito dígitos significativos; após, dígito arredondador, r.

,

6

Operação de multiplicação com o padrão IEEE-754

- 3) arredondamento (continuação)
 - » após acertar expoente e resultado, pode-se finalmente arredondar:
 - ser=0, resultado correto;
 - ullet se r=1 e s=1, soma P+1 para obter o produto dos significandos;
 - se r=1 e s=0, exatamente no meio da faixa IEEE-754 possui quatro modospossíveis, cujo comportamento depende do sinal do resultado.

38

IEEE-754-1985 - Multiplicação, 0 e Precisão

- ${\it extstyle \oslash}$ Se um dos multiplica
ndos é 0, acelera-se multiplicação testando:
 - antes ambos operandos;
 - depois neste caso, cuidado com 0 x ∞, resultado deve ser NaN;
 - sinal deve ser mantido, +0 é diferente de -0;
- Ø Ao multiplicar pode-se precisar do dobro de bits em inteiros; aplicação define se é aceitável apenas a metade inferior do resultado ou se todo o resultado deve ser usado;
- Ø Em linguagens de alto nível, multiplicação inteira sempre usam a primeira opção, mas ponto flutuante é diferente, e as duas são usadas, utilidade é grande na solução de equações lineares;

- Tipicamente, operação em ponto flutuante recebe dois números de mesma precisão e retorna resultado com mesma precisão, p;
- Ø Algoritmo ideal (erro menor) calcula resultado exato e arredonda;
- Multiplicação funciona assim;
- Para soma, existem procedimentos mais efetivos;
- Ø Exemplo com números de 6 bits: 1.10011₂ e 1.10001₂ x 2-5;
- Usando um somador de 6 bits, tem-se: 1.10011

+ 0.00001

1.10100

Operação de adição de a 1 e a 2 - Algoritmo

- 1) Se $e_1 < e_2$, trocar operandos (para que diferença $e_1 e_2 >= 0$). Fazer o expoente do resultado igual a e1, temporariamente;
- 2) Se sinais de a₁ e a₂ diferem, substituir s, por seu complemento de 2;
- 3) Colocar s, em um registrador de p bits e deslocá-lo $d=e_1-e_2$ posições para a direita (entrando com 1's se s, foicomplementado no passo 2). Dos bits deslocados para fora do registrador, guarde o último em um flip-flop g, o penúltimo em um flip-flop r earmazene o ou de todos os restantes como sticky bit.
- 4) ... Este e mais outros quatros passos tão complicados como o passo 3, somando 8 passos no algoritmo. E tudo isso implementado em hardware!!!!

Soma

- passo 4: SOMA
 - » 1.111 + 0.010 = 0.001 e cout=1
- - Se mesmo sinal: desloca para a direita, inserindo 1 e atualiza expoente
 - » 1.0001 e expoente passa a ser 1
- Ø passo 6: atualiza g,r,s
 - -g=1, r=0, s=1
- Ø passo7: arredonda
 - 1.000101 ==> 1.0010 (ou seja: 10.001 = 2,125)
- passo 8: calcula o sinal muito complicado...
 - Se mesmo sinal: não muda

Exercícios

Supor que tenhamos a seguinte representação para ponto flutuante: 1 bit para sinal, 4 para expoente e 7 para a parte fracionária (ou seja, uma norma IEEE 754 para 12 bits, com as mesmas condições para NaN, 0, ±∞)

Qual o valor da polarização e quais os expoentes mínimo e máximo? Valorda polarização: 7.

Expoente máximo +7 (1110-7) e exp mínimo -7 (0000-7).

Converta para base decimal: 0 1010 1011000 +2^(10-7) * 1.1011 = +2^3 * 1.1011 = 1101.1 = 13.5

 $-2^{(7-7)}$ * 1.0101 = 2^0 * 1.0101 = 1.0101 = -1.3125

Ø No exemplo, bit descartado é 1, logo resto deveria ser examinado;

Novamente, apenas se precisa saber se um destes bits é não-zero,

e pode-se assim usar "sticky bits", como na multiplicação;

desde que se guarde o primeiro bit descartado e o "sticky bit" correspondente;

Ø No exemplo acima, o sticky é 1, e o resultado final fica 1.10101₂;

- Subtração é similar, se se trabalha em complemento de dois;
- A seguir, apresenta-se o algoritmo para somar dois números representados no formato IEEE-754-1985.

Soma

- \odot fazer: -1.001 * 2⁻² + -1.111 * 2⁰ (a1 + a2)
- - se a1 < a2 *swap* a1 e a2
 - calcula a distância entre expoentes: (d=2)
 - expoente igual ao expoente máximo (0)
- ø passo2:
 - se sinais diferentes substituir a2 pelo seu complemento
- passo 3: desloca a 2 d dígitos para a direita, seta
 - no exemplo: 0.010 01, logo g=0, r=1, s=0

Supor que tenhamos a seguinte representação para ponto flutuante: 1 bit para sinal, 4 para expoente e 7 para a parte fracionária (ou seja, uma norma IEEE 754 para 12 bits, com as mesmas condições para NaN, $0, \pm \infty$).

Pergunta-se:

- Qual o valor da polarização e quais os expoentes mínimo e máximo?
- Converta os seguintes números para base decimal:
 - 0 1010 1011000 1 0111 0101000
- Multiplicar os dois números acima, mostrando o procedimento da multiplicação para ponto flutuante, com arredondamento
- Qual o resultado, em decimal, arredondado, e qual o erro advindo do arredondamento?

Exercícios

Supor que tenhamos a seguinte representação para ponto flutuante: 1 bit para sinal, 4 para expoente e 7 para a parte fracionária (ou seja, uma norma IEEE 754 para 12 bits, com as mesmas condições para NaN, 0, ±∞).

Multiplicaros dois números anteriores (010101011000 e 1011101010000), $most rando o procedimento da multiplica \\ \tilde{c}a o para ponto flutuante, com$ arredondamento.

```
Resposta: expoente: 10+7-7 = 10
partes fracionárias: 1.1011 * 1.0101 = 10.00110111
fica 2^10 * 10.00110111 = 2^11 * 1.0001101 11
```

Soma-seumaomenossignificativopois round=1 e stick=1 1

```
1 1011 0001 <u>110</u> =
```



Supor que tenhamos a seguinte representação para ponto flutuante: 1 bit para sinal, 4 para expoente e 7 para a parte fracionária (ou seja, uma norma IEEE 754 para 12 bits, com as mesmas condições para NaN, $0, \pm \infty$).

 $Qual\,o\,resultado, em\,decimal, arredondado, e\,qual\,o\,erro\,advindo\,do\,arredondamento?$

Resposta: 17.75, ao invés de 17.71875 (3.5 * 1.3125)

Finalmente acabou (esta parte)!



Fim da Unidade 3!

Aritmética: revisão em 30/novembro/2001 - Moraes e Ney

Aritmética: revisão em 30/novembro/2001 - Moraes e Ney