

Aula 3: Aritmética Computacional - Parte II

Professor(a): Virgínia Fernandes Mota

<http://www.dcc.ufmg.br/~virginiaferm>

OCS (TEORIA) - SETOR DE INFORMÁTICA



- Ponto Flutuante
- Falácias e Armadilhas

- Precisamos de representação para números reais!
- Utilizaremos na representação números normalizados em **notação científica**.
- Notação científica normalizada: possui único dígito à esquerda do ponto decimal (ou ponto binário, caso a base seja 2)

Ex.: $1,0 \times 10^{-9}$

- Ponto flutuante: aritmética computacional que representa os números em que o ponto binário não é fixo.
- Representados na forma $1,aaaaaa \times 2^{bbbb}$
- Precisamos de compromisso entre tamanho da fração e do expoente.
- Aumento da precisão x aumento do intervalo dos números que podem ser representados.

- Números em ponto flutuante múltiplos do tamanho de uma palavra.
 - Precisão simples (floats em C): uma palavra usada na representação
 - Precisão dupla (doubles em C): duas palavras usadas na representação

single: 8 bits
double: 11 bits

single: 23 bits
double: 52 bits

S	Expoente	Fração
---	----------	--------

$$x = (-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$$

- Representações anteriores chamada de **senal e magnitude**
 - Sinal possui um bit separado do restante do número
- Representação em **precisão simples**
 - Números tão pequenos quanto 2×10^{-38} e tão grandes quanto 2×10^{38}
 - Overflow ainda pode ocorrer
 - Expoente positivo torna-se muito grande para caber no campo de expoente
 - Também podemos ter underflow
 - Expoente negativo torna-se muito grande para caber no campo de expoente

- Representação em **precisão dupla**
 - Números tão pequenos quanto 2×10^{-308} e tão grandes quanto 2×10^{308}
- Formato de ponto flutuante do IEEE: **IEEE 754**
- Mais bits podem ser colocados na fração
 - Números sempre na forma 1,xxx
 - Um implícito
 - Bits da fração numerados da esquerda para direita
 - Termo significando utilizado

Ponto Flutuante

Precisão Simples		Precisão Dupla		Objeto Representado
Expoente	Fração	Expoente	Fração	
0	0	0	0	0
0	não zero	0	não zero	\pm número desnormalizado
1-254	qualquer coisa	1-2046	qualquer coisa	\pm número ponto flutuante
255	0	2047		\pm infinito
255	não zero	2047	não zero	NaN (Not a Number)

Codificação de Ponto Flutuante (Números, NaN, Inf)

- Comparação de números seria simplificada se representação do expoente mais negativo fosse próxima a 0000....000 e expoente mais positivo como 1111...111
- $1,0 \times 2^{-1}$: 0 11111111 000000000000000000000000
- $1,0 \times 2^{+1}$: 0 00000001 000000000000000000000000

- Solução: uso de notação deslocada
- Bias adicionado ao expoente
- No padrão IEEE valor 127 utilizado como bias para precisão simples e 1023 para precisão dupla
 $(-1)^s \times (1 + \text{fração}) \times 2^{(\text{expoente} - \text{bias})}$
- No exemplo anterior:
 $-1 + 127 = 126 \rightarrow 0111\ 1110$
 $+1 + 127 = 128 \rightarrow 1000\ 0000$

- Exemplo: Representar $-0,75$ em precisão simples e dupla
- $-0,75_{10} = 0,11_2$
 $0,75 \times 2 = (1),50$
 $0,50 \times 2 = (1),00$
- Na notação científica normalizada: $(-1) \times 1,1 \times 2^{-1}$

- Representando -0,75:
 - $S = 1$
 - Fração = $1000...00_2$
 - Expoente = $-1 + \text{Bias}$
 - Simples: $-1 + 127 = 126 = 01111110_2$
 - Dupla: $-1 + 1023 = 1022 = 01111111110_2$
- Logo:
 - Simples: 1 01111110 1000...00
 - Dupla: 1 01111111110 1000...00

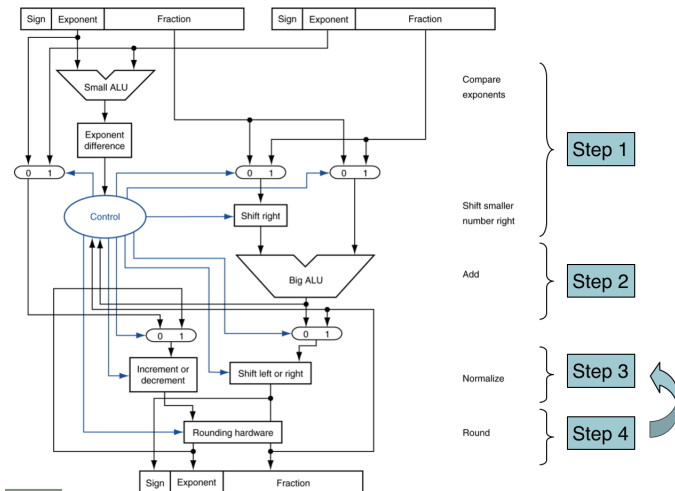
- Que número decimal é representado por este float de precisão simples?

01000000101000...00

- $S = 0$
- Fração = $01000...00_2 = 0 \times 2^{-1} + 1 \times 2^{-2} + \dots$
- Exponente = $10000001_2 = 129$
 $x = (-1)^0 \times (1 + 0,01_2) \times 2^{(129-127)}$
 $= (-1) \times 1,25 \times 2^2$
 $= 5.0$

Ponto Flutuante - Adição

- A adição de ponto flutuante é bem mais complexa que a adição de inteiros.



- Considere a seguinte soma:
 $9,999 \times 10^1 + 1,610 \times 10^{-1}$
- 1. Alinhar as casas decimais: Shift no número de menor expoente
 $9,999 \times 10^1 + 0,016 \times 10^1$
- 2. Somar
 $9,999 \times 10^1 + 0,016 \times 10^1 = 10,015 \times 10^1$
- 3. Normalizar resultado e checar over/underflow
 $1,0015 \times 10^2$
- 4. Arredondar e (re)normalizar se necessário
 $1,002 \times 10^2$

- Instruções de ponto flutuante no MIPS
 - Adição simples e dupla: add.s e add.d
 - Subtração simples e dupla: sub.s e sub.d
 - Multiplicação simples e dupla: mul.s e mul.d
 - Divisão simples e dupla: div.s e div.d
 - Comparação simples e dupla: c.x.s e c.x.d
 - Onde x pode ser igual (eq), diferente (neq), menor que (lt), menor ou igual (le), maior que (gt) ou maior ou igual (ge)
 - Desvio verdadeiro em ponto flutuante (be1t) e falso (bc1f)

- Comparação em ponto flutuante define um bit como verdadeiro ou falso, dependendo da condição de comparação
- Desvio de ponto flutuante decide então se desviará ou não, dependendo da condição
- Projetistas MIPS acrescentaram registradores de ponto flutuante separados: $\$f0, \$f1, \dots, \$f31$
- Usados para precisão simples ou dupla

- Loads e stores separados para ponto flutuante: lwc1 e swc1
- Exemplo: converter código abaixo para assembly

```
1 float f2c (float fahr) {  
2     return ((5.0/9.0) * (fahr - 32.0));  
3 }
```

- Supor que fahr seja passado em \$f12

- Supor que constantes 5.0, 9.0 e 32.0 alcançadas por meio do ponteiro global \$gp e retorno em \$f0

f2c:

```
lwc1 $f16, const5($gp) # $f16 = 5.0
```

```
lwc1 $f18, const9($gp) # $f18 = 9.0
```

```
div.s $f16, $f16, $f18 # $f16 = 5.0/9.0
```

```
lwc1 $f18, const32($gp) # $f18 = 32.0
```

```
sub.s $f18, $f12, $f18 # fahr - 32.0
```

```
mul.s $f0, $f16, $f18 # mult. result. intermediarios
```

```
jr $ra # retorna
```

- Números em PF normalmente são aproximações para um número
- Arredondamento pode ser crítico
- IEEE 754 define dois bits extras para arredondamento durante operações: Guarda e arredondamento

- Exemplo: Somar $2,56 \times 10^0$ a $2,34 \times 10^2$, supondo 3 dígitos significativos
- Guarda e arredondamento garantem 5 bits durante operação
 $2,3400 + 0,0256 = 2,3656 = 2,37$
- Sem guarda e arredondamento teríamos:
 $2,34 + 0,02 = 2,36$

- Falácia: Adição de ponto flutuante é associativa

$$x = -1,5 \times 10^{38}, y = 1,5 \times 10^{38} \text{ e } z = 1,0$$

$$x + (y + z) = -1,5 \times 10^{38} + (1,5 \times 10^{38} + 1,0) = -1,5 \times 10^{38} + 1,5 \times 10^{38} = 0,0$$

$$(x + y) + z = (-1,5 \times 10^{38} + 1,5 \times 10^{38}) + 1,0 = 0,0 + 1,0 = 1,0$$

- Falácia: Deslocamento à direita é o mesmo que uma divisão de inteiros por uma potência de dois
 - Verdade APENAS para inteiros sem sinal
- 5: 1111 1111 1111 1111 1111 111 1111 1011
- Dois deslocamentos à direita não é o mesmo que dividir por 4*
- 0011 1111 1111 1111 1111 1111 1111 1110
- 1.073.741.822 e não -1

- Falácia: Somente matemáticos teóricos se importam com precisão do ponto flutuante
Intel Pentium FDIV bug
Erros no excel
Outros??
Pesquisar!!!!

- Converta os números decimais em sua forma binária (em complemento a 2):
 - a) -2
 - b) 10
 - c) 7550
 - d) 13,25
 - e) -0,4217
- Converta os números binários em sua forma decimal:
 - a) $(10100)_2$
 - b) $(1101)_2$
 - c) $(0,1101)_2$
 - d) $(11101,01)_2$
- Coloque os números no padrão IEEE 754

Exercícios e Prova

