Dolnośląski Festiwal Nauki na Politechnice Wrocławskiej

Beata Laszkiewicz i Krystyna Ziętak

Co każdy powinien wiedzieć o obliczeniach liczbowych w komputerze?

Instytut Matematyki
Wydział Podstawowych Problemów Techniki
21 września 2004

Kto to jest **William Kahan**?

Zajrzyj na jego stronę:

http://www.cs.berkeley.edu/~wkahan/

Plan wykładu

1. Postać zmiennopozycyjna liczby dziesiętnej.
2. System dwójkowy: zaokrąglenie i obcięcie.
3. *IEEE Standard*, błąd reprezentacji zmiennopozycyjnej liczby, liczby typu: *single i double.*
4. **Prezentacja komputerowa**:
   Jak pamięta się liczbę w komputerze?
5. Arytmetyka zmiennopozycyjna, utrata cyfr znaczących.
6. **Prezentacja komputerowa**:
   Jak obliczyć $a^2 - b^2$*?*
7. **Prezentacja komputerowa**:
   Czy dodawanie jest przemienne?
8. **Prezentacja komputerowa**: *Algorytmy obliczania wartości wielomianu* $(x-1)^8$.
9. Czy w banku dokładnie obliczają odsetki?
10. Dzielenie przez zero i NaNy.

11. **Prezentacja komputerowa**:
    Czy wiesz, jak obliczyć pierwiastki trójmianu kwadratowego?
12. **Prezentacja komputerowa**:
    Czy wracając zawsze trafisz w punkt wyjścia, czyli o obliczaniu w komputerze wyrazów pewnego ciągu.
13. Numeryczne obliczanie pochodnej, czyli czy warto dążyć do zera?
14. **Prezentacja komputerowa**: *Obliczanie przybliżonej wartości pochodnej.*
15. Analiza dwóch algorytmów obliczania w komputerze $c = a^2 - b^2$.
16. Liczby w komputerze - podsumowanie
17. Odpowiedź na pytanie:
    Kto to jest Kahan?

Liczby

- liczby całkowite: $0, 1, -1, 2, -2, \ldots$

- liczby wymierne: $\frac{1}{2} = 0.5,\ \frac{4}{3} = 1.333\ldots$

- liczby niewymierne:

$$\sqrt{2} = 1.41421\ldots$$

$$\pi = 3.14159\ldots$$

$$e = \lim_{n\to\infty} (1 + \frac{1}{n})^n = 2.718281845\ldots$$

liczba Eulera

- liczby rzeczywiste

System dziesiętny

liczby całkowite

$$102 = 1 \times 10^2 + 0 \times 10^1 + 2 \times 10^0 = (102)_{10}$$

$$x = x_k \times 10^k + x_{k-1} \times 10^{k-1} + \ldots + x_0 \times 10^0$$

$$x = (x_k x_{k-1} \ldots x_0)_{10}$$

$$x_i \in \{0, 1, \ldots, 9\}$$

Liczby rzeczywiste

$$102 = 10.2 \times 10^1 = 1.02 \times 10^2 = 0.102 \times 10^3$$

$$0.0102 = 0.102 \times 10^{-1}$$

$$0.102 = 1 \times 10^{-1} + 0 \times 10^{-2} + 2 \times 10^{-3}$$

Postać zmiennopozycyjna dziesiętna

$$x = m10^c = mEc$$

m - **mantysa**
c - **cecha (wykładnik)**

$$m = 0.m_1m_2 \cdots m_t$$

$$m_1 \neq 0, \quad 0.1 \leqslant m < 1$$

$$102 = 0.102E3$$

$$m = 0.102, \quad c = 3$$

$$0.0102 = 0.102E - 1$$

$$m = 0.102, \quad c = -1$$

System dwójkowy (binarny)

podstawa systemu: 2

cyfry: 0, 1

$$x = \frac{11}{2} = \frac{11}{16} \times 8 =$$

$$= 0.6875 \times 2^3$$

mantysa: $m = \frac{11}{16} = \frac{1}{2} + \frac{1}{8} + \frac{1}{16}$

$$m = (0.6875)_{10} = (0.1011)_2$$

cecha: $c = 3 = 2 + 1$

$$c = (11)_2$$

$$\boxed{x = 0.1011E11}$$

$$\boxed{x = 0.m_1 m_2 \cdots m_t \; E \; c}$$

$$m = 0.m_1 m_2 \cdots m_t$$

$$m_1 = 1, \qquad 1/2 \leqslant m < 1$$

$$m = m_1 \times \frac{1}{2} + m_2 \times \frac{1}{4} + m_3 \times \frac{1}{8} + \ldots + m_t \times \frac{1}{2^t}$$

$$m_i \in \{0, 1\}$$

$$x = (\frac{m_1}{2} + \frac{m_2}{4} + \frac{m_3}{8} + \ldots + \frac{m_t}{2^t}) \times 2^c$$

Zaokrąglenie i obcięcie

$$x = \frac{1}{10} = \frac{8}{10} \times \frac{1}{8}$$

$$m = \frac{4}{5}, \qquad c = -3$$

$$m = 0.1100110011001100\cdots$$

$m = 0.110011001100$ 12 cyfr, **obcięcie**
$m = 0.110011001101$ 12 cyfr, **zaokrąglenie**

$$\mathbf{m} = (0.1100)_2 = (\mathbf{0.75})_{10} \qquad \textbf{obcięcie}$$

$$\tilde{\mathbf{x}} = 0.75 \times 2^{-3} = \frac{\mathbf{3}}{\mathbf{32}} = \mathbf{0.093705}$$

$$\mathbf{m} = (0.1101)_2 = (\mathbf{0.8125})_{10} \qquad \textbf{zaokr.}$$

$$\tilde{\mathbf{x}} = 0.8125 \times 2^{-3} = \frac{\mathbf{13}}{\mathbf{128}} = \mathbf{0.101565}$$

System zmiennopozycyjny dwójkowy

Przykład

$$t = 3, \quad c_{\min} = -1, \quad c_{\max} = 3$$

mantysy: $0.100 \quad 0.101 \quad 0.110 \quad 0.111$

cechy: $-1, \quad 0, \quad 1, \quad 2, \quad 3$

nieujemne liczby zmiennopozycyjne:

- $0, \quad 0.25 \quad 0.3125 \quad 0.3750 \quad 0.4375$
- $0.5 \quad 0.625 \quad 0.75 \quad 0.875$
- $1.0, \quad 1.25 \quad 1.5 \quad 1.75$
- $2.0, \quad 2.5, \quad 3.0 \quad 3.5$
- $4.0 \quad 5.0 \quad 6.0 \quad 7.0$

IEEE Standard - 1985 rok

Floating-point numbers and roundoff errors

$$x = m \times 2^c$$

$x > 0$, liczba rzeczywista

m mantysa, c cecha

$$c_{\min} \leqslant c \leqslant c_{\max}$$

Uwaga: Teraz $1 \leqslant m < 2$

$$m = (m_0.m_1 m_2 \cdots m_{t-1})_2$$

$$m_0 = 1, \qquad m_1, \ldots, m_{t-1} \in \{0, 1\}$$

t - liczba cyfr (binarnych) mantysy

znak	cecha	mantysa

Mantysa t-bitowa:

$$m = 1.m_1 m_2 \ldots m_{t-1} = 1.f$$

$$f = m_1 \ldots m_{t-1}$$

Uwagi:

- Nie zapamiętuje się jawnie w komputerze bitu m_0, bo on jest zawsze równy 1.

- Cechę pamięta się w innej postaci (przesuniętej): $\tilde{c} = c + bias$, gdzie $bias$ taki, że $\tilde{c} > 0$

- **epsilon maszynowy** $\epsilon_M = 2^{1-t}$
  odległość liczby 1 od najbliższej liczby zmiennopozycyjnej w komputerze większej niż 1

$$1 + \epsilon_M > 1$$

ulp - **unit in last place**

Błąd reprezentacji liczby w zmiennopozycyjnej arytmetyce dwójkowej

$$\mathrm{fl}(x) = \pm(m_0.m_1m_2\ldots m_{t-1})_2 \times 2^c$$

$$\delta = \frac{\mathrm{fl}(x) - x}{x}$$

δ - błąd względny reprezentacji zmiennopozycyjnej liczby x

$$\boxed{\mathrm{fl}(x) = x(1+\delta), \qquad |\delta| \leqslant u}$$

$\boxed{u = 2^{-t}}$ **unit roundoff**

- **single precision - 32 bity**
  mantysa ze znakiem t=23+1 bity
  cecha 8 bitów

  $c_{\min} = -126, \quad c_{\max} = 127, \quad bias = 127$

  zakres liczb: $10^{\pm 38}$

  precyzja obliczeń:

  $$\mathbf{u = 2^{-24} \approx 5.96 \times 10^{-8}}$$

- **double precision - 64 bity**
  mantysa $t = 52 + 1$, cecha 11 bitów
  $c_{\min} = -1022, \quad c_{\max} = 1023$

  zakres liczb: $10^{\pm 308}$

  precyzja obliczeń:

  $$\mathbf{u = 2^{-53} \approx 1.11 \times 10^{-16}}$$

Przykład: **typ double** (64 bity)

mantysa $52+1$ bity, cecha 11 bitów, $bias = 1023$

$$x = 0.09375 = \frac{3}{32} = \frac{6}{4} \times 2^{-4} =$$

$$= (1 + \frac{1}{2}) \times 2^{-4} = (1.10\cdots0)_2 \times 2^{-4}$$

- **mantysa**: znak: 0
  faktyczna matysa $m = 1.10\cdots0$
  pamiętana w komputerze mantysa $10\cdots0$
- **cecha**: $c = -4$, $\tilde{c} = c + bias = 1019$

$$1019 = 512{+}256{+}128{+}64{+}32{+}16{+}8{+}2{+}1$$

$$\tilde{c} = (01111111011)_2$$

$$\boxed{0|1000\cdots0|01111111011}$$

Prezentacja komputerowa

Jak pamięta się liczbę w komputerze?

$$x = m \times 2^c$$

znak	cecha	mantysa

Utrata cyfr znaczących

$$x = 0.372\underline{\mathbf{1478693}}, \qquad y = 0.372\underline{\mathbf{0230572}}$$

$$x - y = 0.0001248121 = 0.1248121 \times 10^{-3}$$

$$x - y \quad \text{po zoakragleniu} \quad \underline{\mathbf{0.12481}} \times 10^{-3}$$

pięć cyfr znaczacych

$$\mathrm{fl}(x) = 0.37215, \qquad \mathrm{fl}(y) = 0.37202$$

$$\mathrm{fl}(x) - \mathrm{fl}(y) = 0.00013 = \underline{\mathbf{0.13000}} \times 10^{-3}$$

$$\left|\frac{x - y - [\mathrm{fl}(x) - \mathrm{fl}(y)]}{x - y}\right| \approx 0.04$$

Arytmetyka zmiennopozycyjna

$$\mathrm{fl}(\mathbf{a}) = \mathbf{a}, \quad \mathrm{fl}(\mathbf{b}) = \mathbf{b}$$

$$\mathrm{fl}(\mathbf{a} \pm \mathbf{b}) = (\mathbf{a} \pm \mathbf{b})(1 + \delta_1)$$

$$\mathrm{fl}(\mathbf{a} \times \mathbf{b}) = (\mathbf{a} \times \mathbf{b})(1 + \delta_2)$$

$$\mathrm{fl}(\mathbf{a}/\mathbf{b}) = (\mathbf{a}/\mathbf{b})(1 + \delta_3)$$

$$|\delta_i| \leqslant u = 2^{-t}$$

Zasada w IEEE:

każde działanie arytmetyczne daje wynik

dzielenie przez zero!

Prezentacja komputerowa

Jak obliczać $\mathbf{a^2 - b^2}$?

$$\mathbf{c = (a - b)(a + b) = a^2 - b^2}$$

$$\frac{1 + \frac{b^2}{a^2}}{|1 - \frac{b^2}{a^2}|}, \qquad \frac{b^2}{a^2} \approx 1$$

Prezentacja komputerowa

Czy dodawanie jest przemienne?

$$\mathbf{d} = \mathrm{fl}((\mathbf{a} + \mathbf{b}) + \mathbf{c}) \neq \mathrm{fl}(\mathbf{a} + (\mathbf{b} + \mathbf{c}))$$

Prezentacja komputerowa

Algorytmy obliczania

wartości wielomianu $w(x) = (x-1)^8$

wykres $w(x)$ na przedziale $[0.99,\ 1.01]$

- $a_8x^8 + a_7x^7 + \cdots + a_1x + a_0$
- schemat Hornera:
  $(\cdots((a_8x + a_7)x + a_6)x + \cdots)$
- $(((x-1)^2)^2)^2$
- $\mathrm{e}^{(8\ \ln\ (\mathrm{abs}\ (x-1)))}, \qquad x \neq 1$

Oszczędności w banku

5 procent za rok, $\quad r = 0.05$

$$a_1 = a_0 \times (1 + 0.05/4)$$

po pierwszym kwartale

$$a_2 = a_1 \times (1 + 0.05/4) = a_0(1 + 0.05/4)^2$$
$$a_3 = a_2 \times (1 + 0.05/4) = a_0(1 + 0.05/4)^3$$
$$a_4 = a_3 \times (1 + 0.05/4) = a_0(1 + 0.05/4)^4$$

$$\boxed{f = a_0 \times (1 + \tfrac{r}{n})^n}$$

$$z = x^y = \mathrm{e}^{(y \ \ln \ x)}$$

$$\boxed{f = a_0 \times \mathrm{e}^{n \ \ln \ (1+r/n)}}$$

single precision C
M. Overton

n	koncowe konto I	koncowe konto II
1	$1.050000E+02$	$1.050000E+02$
4	$1.050945E+02$	$1.050945E+02$
365	$1.0512\underline{\mathbf{68}}E+02$	$1.0512\underline{\mathbf{67}}E+02$
10000	$1.0512\underline{\mathbf{94}}E+02$	$1.0512\underline{\mathbf{71}}E+02$
20000	$1.0512\underline{\mathbf{02}}E+02$	$1.0512\underline{\mathbf{71}}E+02$

$$f = a_0 \times (1 + \frac{r}{n})^n$$

$$f = a_0 \times \mathrm{e}^{n \ \ln \ (1+r/n)}$$

$$a_0 = 100 \ USD$$

Dzielenie przez zero i NaNy

Zasada w IEEE:

każde działanie arytmetyczne daje wynik

	$przyklad$	$wynik$
$invalid\ operation$	$0/0$	NaN
$invalid\ operation$	$0 \times \infty$	NaN
$invalid\ operation$	∞/∞	NaN
$invalid\ operation$	$\infty - \infty$	NaN
$invalid\ operation$	$\sqrt{-1}$	NaN
$overflow$		$\pm\infty$
$underflow$		$subnormal$
$divide\ by\ zero$	$x/0$	$\pm\infty$

NaN: Not a Number

specjalny ciąg bitów
mantysa niezerowa
specjalna cecha

Prezentacja komputerowa

Czy wiesz, jak obliczać pierwiastki trójmianu kwadratowego?

$$x^2 + px + q = 0$$

$$\Delta = p^2 - 4q$$

$$x_1 = \begin{cases} \frac{-p+\sqrt{\Delta}}{2} & \text{jesli } p \leqslant 0 \\ \\ \frac{2q}{-p-\sqrt{\Delta}} & \text{jesli } p > 0 \end{cases}$$

$$x_1 = \frac{-p+\sqrt{\Delta}}{2} = \frac{(-p+\sqrt{\Delta})(-p-\sqrt{\Delta})}{2(-p-\sqrt{\Delta})} = \frac{q}{x_2}$$

$$x_2 = \frac{-p-\sqrt{\Delta}}{2}$$

Prezentacja komputerowa

Czy wracając zawsze trafisz w punkt wyjścia, czyli o obliczaniu w komputerze wyrazów pewnego ciągu

- a - dany parametr całkowity
- $x_1 = 1, \quad x_2 = 1/a$
- $ax_{i-1} - 10x_i + ax_{i+1} = 0$
- $x_{i+1} = \frac{10}{a}x_i - x_{i-1}$
- $x_{i-1} = \frac{10}{a}x_i - x_{i+1}$

Numeryczne obliczanie pochodnej, czyli czy warto dążyć do zera

$$f'(a) \approx \frac{f(a+h) - f(a)}{h} \equiv p(a,h)$$

$$f(a) = \sin\,(a), \qquad f'(a) = \cos\,(a)$$

$$blad = p(a,h) - f'(a)$$

$\cos\,(1) = 5.403023E{-}01$, double precision C++

h	$p(1,h)$	$blad$
$1.0E-02$	$5.360860E-01$	$-4.216325E-03$
$1.0E-07$	$5.403023E-01$	$-4.182769E-08$
$1.0E-08$	$5.403023E-01$	$-2.969885E-09$
$1.0E-09$	$5.403024E-01$	$+5.254127E-08$
$1.0E-10$	$5.403022E-01$	$-5.848104E-08$
$1.0E-11$	$5.403011E-01$	$-1.1668704E-06$
$1.0E-15$	$5.551115E-01$	$+1.480921E-02$
$1.0E-20$	$0.000000E+00$	$-5.4403023E-01$

Prezentacja komputerowa

Obliczanie przybliżonej wartości pochodnej

a - ustalone

$h = 10^{-x}, \quad x$ jest zmieniane, $\quad h \to 0$

$$f'(a) \approx \frac{f(a+h) - f(a)}{h} \equiv p(a,h)$$

$$blad = p(a,h) - f'(a)$$

Skala logarytmiczna

na osi $0x$ jest $x = -\log_{10} h$

na osi $0y$ jest $y = -\log_{10} |blad|$

Przykład

$h = 10^{-2}, \qquad x = -\log_{10} h = 2$

Im większe x, tym mniejsze jest h.

$|blad| = 10^{-14}, \qquad y = -\log_{10} |blad| = 14$

Im większe y tym jest mniejszy $|blad|$

Analiza błędów zaokrągleń

$$\mathbf{c} = (\mathbf{a} - \mathbf{b})(\mathbf{a} + \mathbf{b}) = \mathbf{a}^2 - \mathbf{b}^2$$

Algorytm I

$$\begin{gathered}
\mathrm{fl}((\mathbf{a} - \mathbf{b})(\mathbf{a} + \mathbf{b})) = \\
= (\mathbf{a} - \mathbf{b})(1 + \delta_1)(\mathbf{a} + \mathbf{b})(1 + \delta_2)(1 + \delta_3) = \\
= (\mathbf{a}^2 - \mathbf{b}^2)(1 + \beta) \\
|\delta_i| \leqslant u = 2^{-t}
\end{gathered}$$

$$1+\beta = 1+\delta_1+\delta_2+\delta_3+\delta_1\delta_2+\delta_1\delta_3+\delta_2\delta_3+\delta_1\delta_2\delta_3$$

$$|\beta| \leqslant 3u = 3 \times 2^{-t}$$

Algorytm II

$$\mathrm{fl}(\mathbf{a}^2 - \mathbf{b}^2) =$$
$$= [(\mathbf{a} \times \mathbf{a})(1+\delta_1) - (\mathbf{b} \times \mathbf{b})(1+\delta_2)](1+\delta_3) =$$
$$= (\mathbf{a}^2 - \mathbf{b}^2)(1+\beta)$$

$$1 + \beta = (1 + \frac{\delta_1 \mathbf{a}^2 - \delta_2 \mathbf{b}^2}{\mathbf{a}^2 - \mathbf{b}^2})(1+\delta_3)$$

$$\frac{|\delta_1 \mathbf{a}^2 - \delta_2 \mathbf{b}^2|}{|\mathbf{a}^2 - \mathbf{b}^2|} \leqslant \frac{|\delta_1|\mathbf{a}^2 + |\delta_2|\mathbf{b}^2}{|\mathbf{a}^2 - \mathbf{b}^2|} \leqslant$$
$$\leqslant \frac{\mathbf{a}^2 + \mathbf{b}^2}{|\mathbf{a}^2 - \mathbf{b}^2|} \times 2^{-t} = \frac{1 + \mathbf{b}^2/\mathbf{a}^2}{|1 - \mathbf{b}^2/\mathbf{a}^2|} \times 2^{-t}$$

Liczby - podsumowanie

typ double: $x = 1.f \times 2^c$

- mantysa $52 + 1$ bitów
- cecha 11 bitów
- $bias = 1023$
- pamiętana cecha: $\tilde{c} = c + bias$

$$x = \begin{cases} \pm(1.f) \times 2^{\tilde{c}-1023} & \text{jesli } 0 < \tilde{c} < 2047 \\ \pm 0 & \text{jesli } \tilde{c} = 0, f = 0 \\ \pm\infty & \text{jesli } \tilde{c} = 2047, f = 0 \\ \mathbf{NaN} & \text{jesli } \tilde{c} = 2047, f \neq 0 \\ \pm(0.f) \times 2^{-1022} & \text{jesli } \tilde{c} = 0, f \neq 0 \end{cases}$$

denormalized (subnormal) numbers

- **if** $z \neq 0$ **then** $s := 3/z$ **else** $s := 3$

- **if** $1 \times z \neq 0$ **then** $s := 3/z$ **else** $s := 3$

- **if** $1 + |z| \neq 1$ **then** $s := 3/z$ **else** $s := 3$

- $x := y + z$

  if $x \neq y + z$ **then** *print*"'*why not*?"

Więcej zob. W. Kahan:

Why do we need a floating-point arithmetic standard?
Berkeley 1981 (druga edycja 2001).

Kto to jest **William Kahan**?

Odpowiedź

Laureat nagrody imienienia Turinga
przyznanej w roku 1989
przez ACM za IEEE standard

ACM:
Association for Computing Machinery

IEEE:
Institute of Electrical and Electronics Engineers, New York, USA

---

http://www.cs.berkeley.edu/~wkahan/