Kod IEEE754

IEEE – Institute of Electrical and Electronics Engineers

IEEE754 (1985) - norma dotycząca zapisu binarnego liczb zmiennopozycyjnych (pojedynczej precyzji)

Liczbę binarną o postaci

$$(-1)^{s} \cdot 1.f \cdot 2^{e-127}$$

zapisuje się na 32-bitach następująco:

IEEE754 c.d.

$$(-1)^{s} \cdot 1.f \cdot 2^{e-127}$$

Mantysa zapisywana jest bez wiodącej cyfry 1, co pozwala oszczędzić miejsce

Wykładnik zapisany jest w formie przesuniętej, tj. zwiększony o 127, co pozwala na uniknięcie liczb ujemnych jego zapisie.

$$1100000010100000\ldots000_b =$$

$$-1 \cdot 1.01_b \cdot 2^{129-127} = -1.01_b \cdot 2^2 = -5$$

$$0011000010000000\ldots000_b =$$

$$+1 \cdot 1.0_b \cdot 2^{96-127} = 1 \cdot 2^{-31} \approx 4.65e{-}10$$

Dec → IEEE754

-7.25

↓

-111.01_b

↓

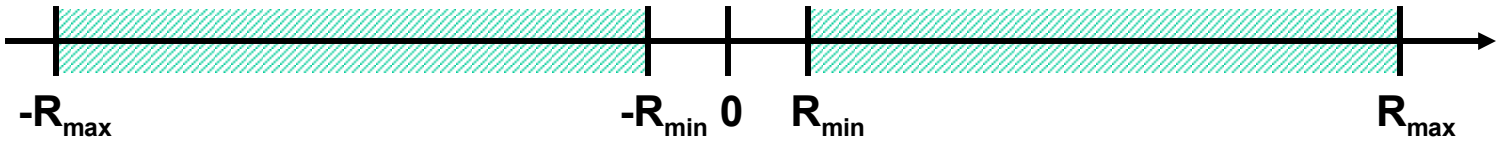
$-1.1101_b \cdot 2^2$

↓

1 10000001 11010000000000000000000

Ograniczenia zapisu zmiennopozycyjnego

Ograniczenie zakresu:
skończona długość pola wykładnika



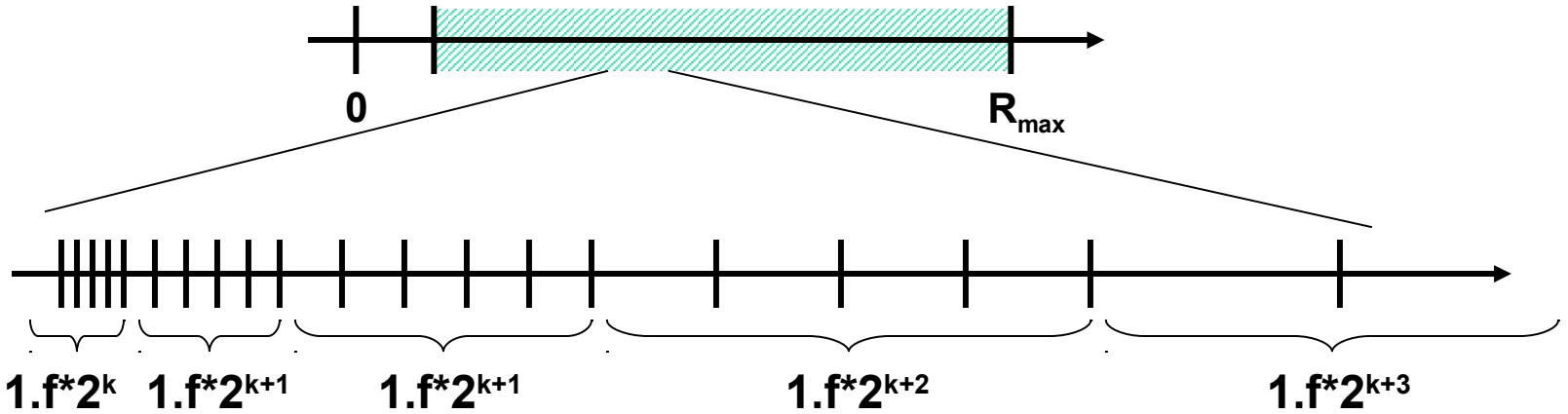
IEEE754 single precision (32 bity)

$R_{min} \approx 1.2e{-}38$

$R_{max} \approx 3.4e{+}38$

Ograniczenia zapisu zmiennopozycyjnego

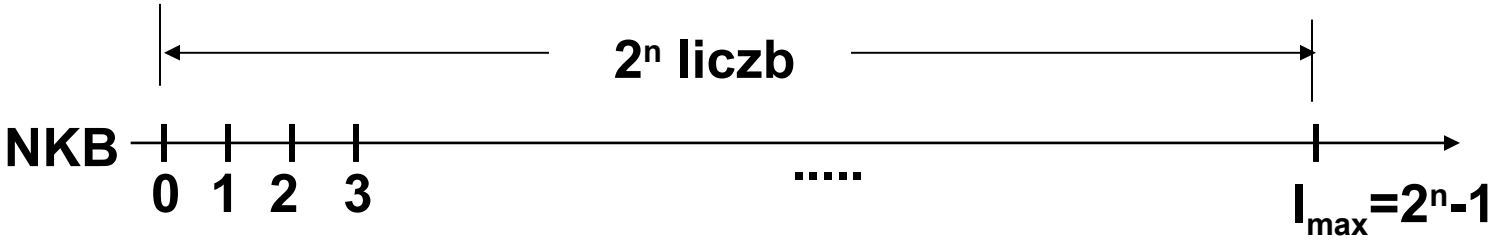
Ograniczenie dokładności: skończona długość pola mantysy



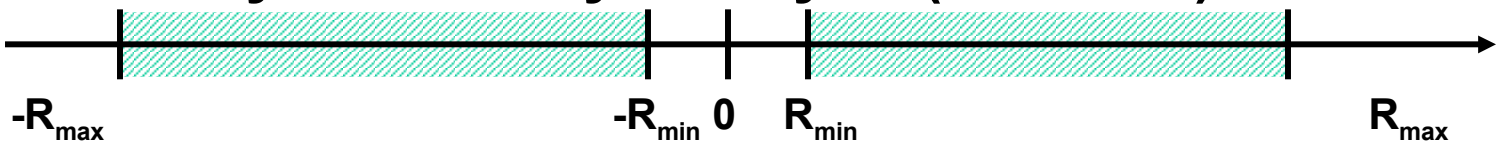
Zapisać można tylko niektóre liczby wymierne.
„Zagęszczenie” liczb jest zmienne i zależy od wartości wykładnika.
W każdym przedziale pomiędzy 2^{i} i 2^{i+1} znajduje się tyle równomiernie rozłożonych liczb, na ile kombinacji pozwala długość pola mantysy.
Dla liczb bliskich R_{min} dokładność jest największa, ale zakres najmniejszy, dla liczb bliskich R_{max} – najmniejsza, a zakres największy.

Liczby całkowite vs zmiennopozycyjne

Za pomocą n-bitów można zapisać dokładnie 2^n różnych liczb całkowitych (NKB, U2)



Za pomocą n-bitów można zapisać mniej niż 2^n różnych liczb wymiernych (IEEE754)



Z n-bitów można utworzyć 2^n różnych kombinacji binarnych.
Znaczenie tych kombinacji zależy od interpretacji.
W przypadku zapisu zmiennopozycyjnego, dostępne wartości są jedynie inaczej rozłożone na osi liczbowej, ale jest ich niemal tyle samo co liczb całkowitych.

Single vs Double Precision IEEE754

Single Precision: 32 bity

8b wykładnik + 23b mantysa

$R_{min} \approx 10^{-38}$

$R_{max} \approx 10^{+38}$

dokładność około 7 cyfr znaczących

Double Precision: 64 bity

11b wykładnik + 52b mantysa

$R_{min} \approx 10^{-308}$

$R_{max} \approx 10^{+308}$

dokładność około 16 cyfr znaczących

Arytmetyka liczb zmiennopozycyjnych

Dla zapisu zmiennopozycyjnego IEEE754:

1. **Niemożliwe jest zapisanie wszystkich liczb z dostępnego zakresu.**
2. **Działania arytmetyczne (+,-,*,/) dają wyniki obarczone błędem przybliżenia.**
3. **Błąd zakresu (*niedomiar* lub *przepełnienie*) jest sygnalizowany kodem specjalnym IEEE754**
4. **Błąd przybliżenia nie jest sygnalizowany.**
5. **Operacje arytmetyczne wymagają skomplikowanych algorytmów.**

Kody specjalne IEEE754

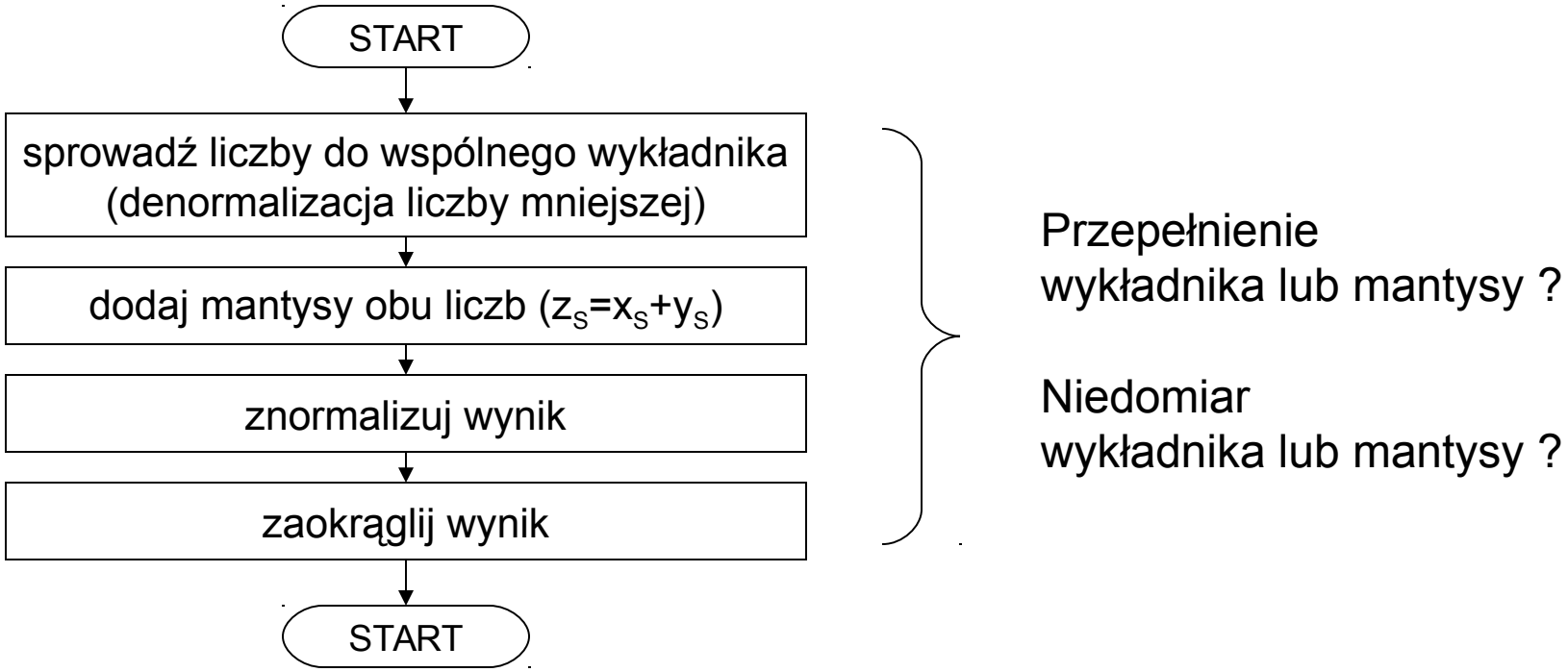
	znak	wykładnik	mantysa
liczba dodatnia	0	1-254	mantysa
liczba ujemna	1	1-254	mantysa
liczba zero+ (0+)	0	0	0
liczba zero- (0-)	1	0	0
liczba zdenormalizowana	0/1	0	mantysa
+nieskończoność	0	255	0
-nieskończoność	1	255	0
NaN (Not a Number)	0/1	255	≠0 (kod błędu)

Dodawanie liczb zmiennopozycyjnych (*fp*)

$$x = x_S \cdot 2^{x_E}$$

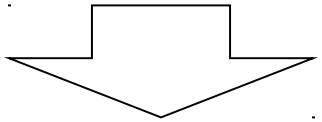
$$y = y_S \cdot 2^{y_E}$$

$$\Rightarrow \quad z = x + y \quad \Rightarrow \quad z = z_S \cdot 2^{z_E}$$



Sprowadzanie do wspólnego wykładnika - denormalizacja

$$1.000000 \; 2^{+4} + 1.000000 \; 2^{-4}$$



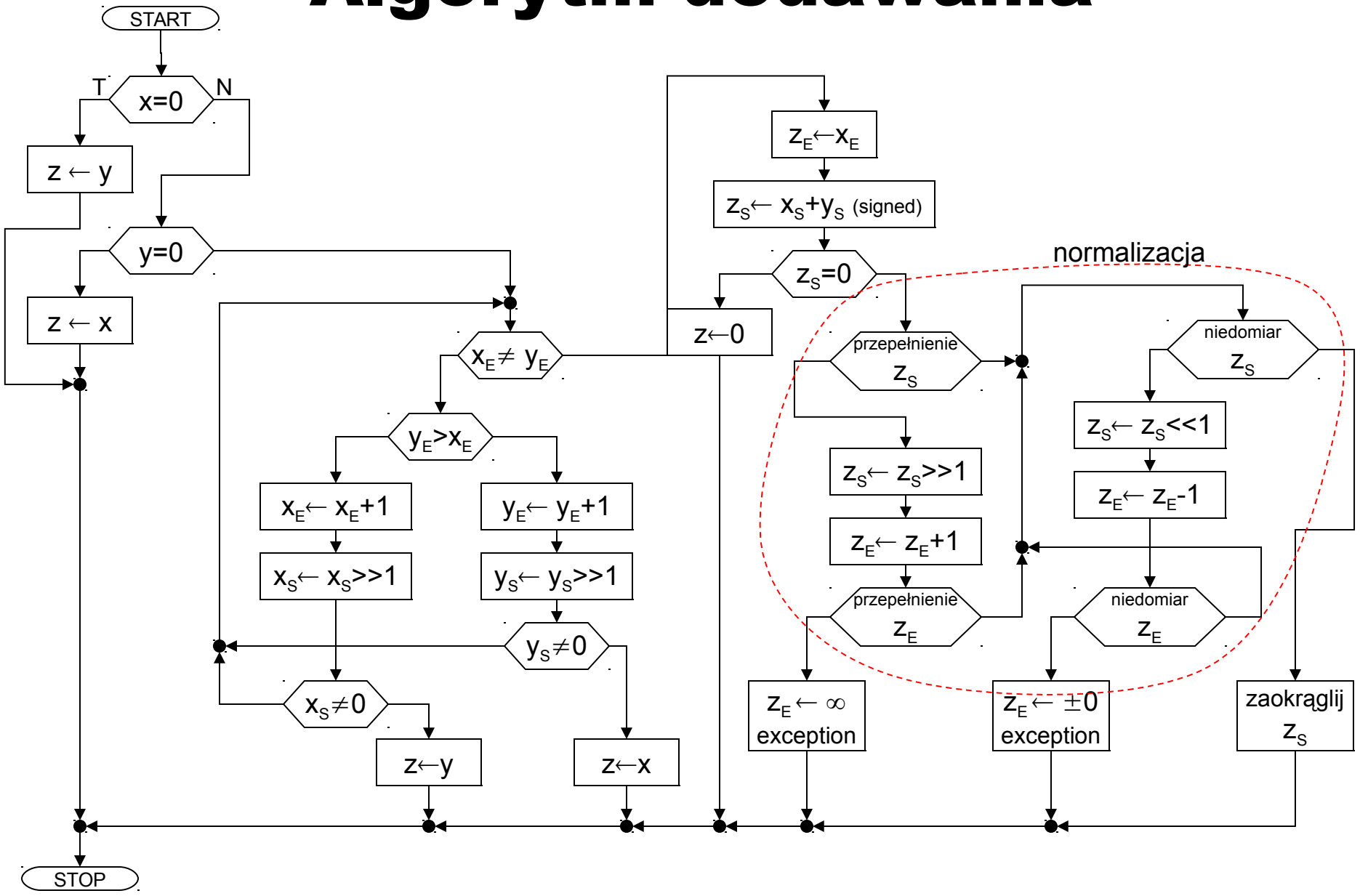
$$1.000000 \; 2^{+4} + 0.000000\,01 \; 2^{+4}$$

Denormalizacja: przesunięcie mantysy w prawo + zwiększanie wykładnika

!

W przypadku dodawania liczb znacznie różniących się wartością wykładnika liczba mniejsza jest tracona (nie ma wpływu na wynik dodawania)

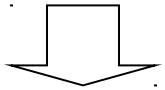
Algorytm dodawania



Zaokrąglanie mantysy

zaokrąglanie w „górę”

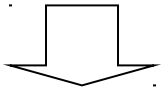
z_S = ... 00 | 1..01



z_S = ... 01 |

zaokrąglanie w „dół”

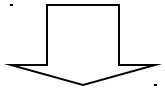
z_S = ... 00 | 0..01



z_S = ... 00 |

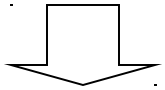
zaokrąglanie do najbliższej parzystej

z_S = ... 00 | 100..



z_S = ... 00 |

z_S = ... 01 | 100..

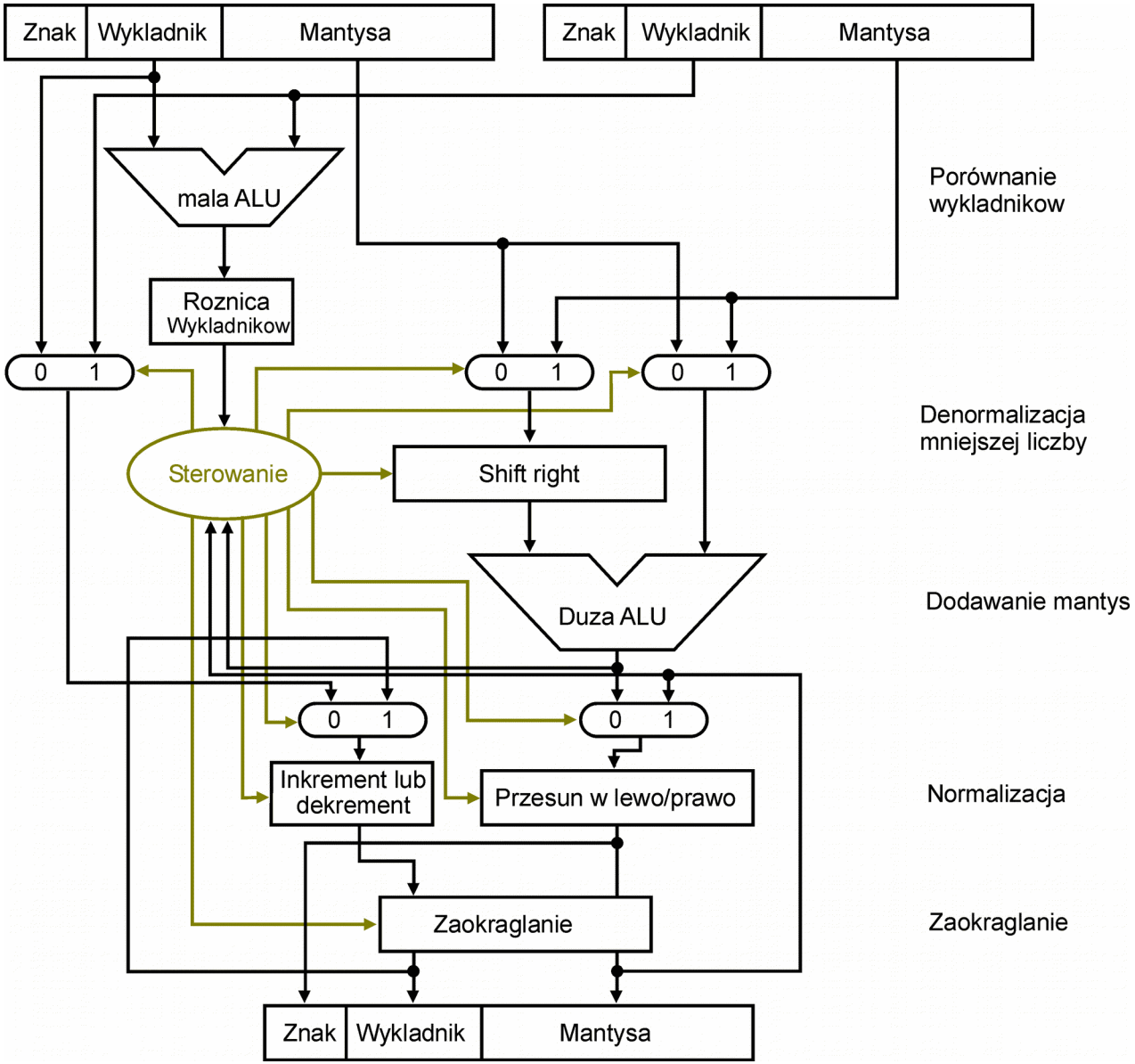


z_S = ... 10 |

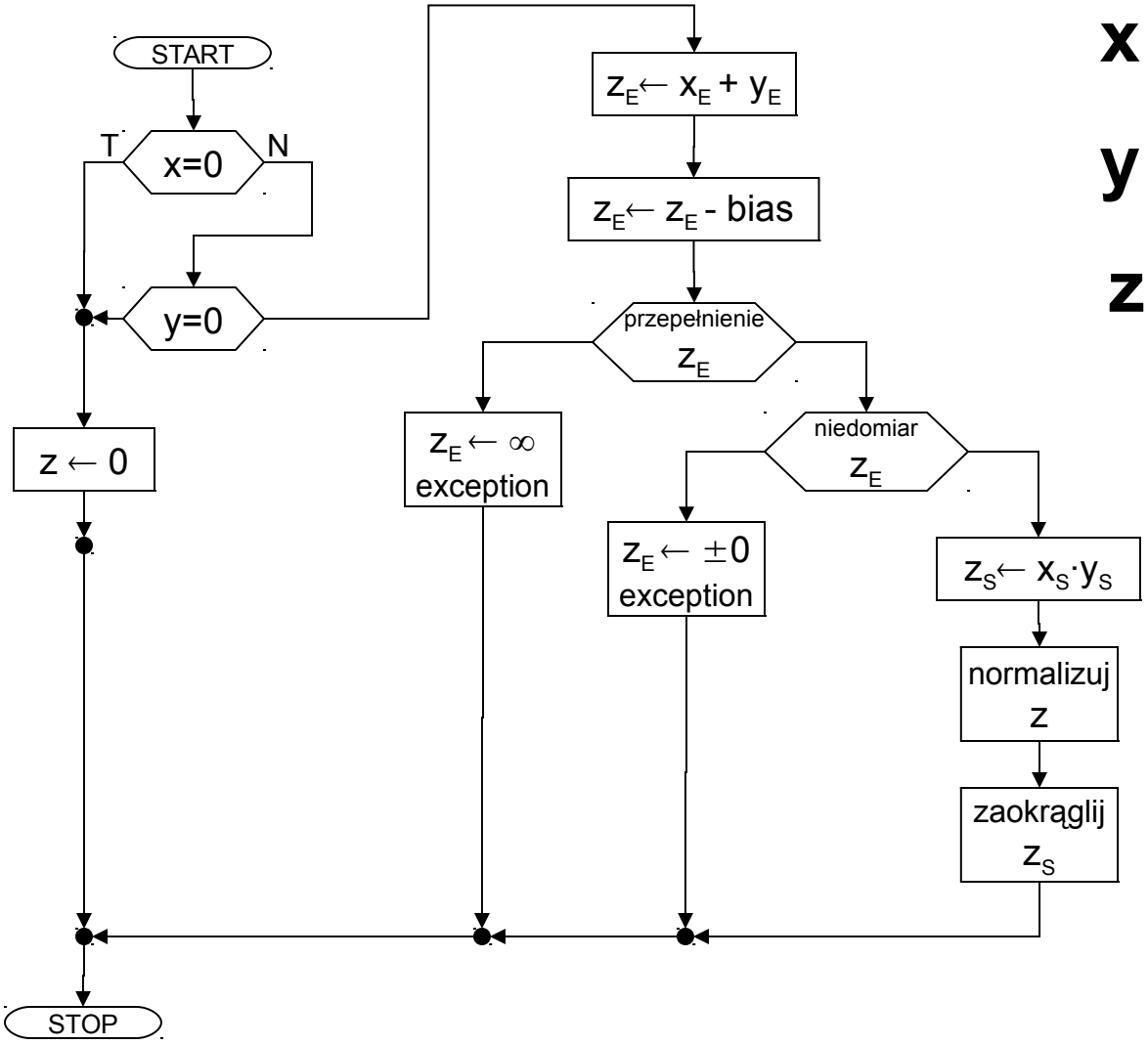
Rejestry do wykonywania działań na mantysach muszą być dłuższe od docelowego rozmiaru mantysy

Przyjęte reguły zaokrąglania umożliwiają otrzymywanie deterministycznych (powtarzalnych) wyników.

Dodawanie - Hardware



Algorytm mnożenia

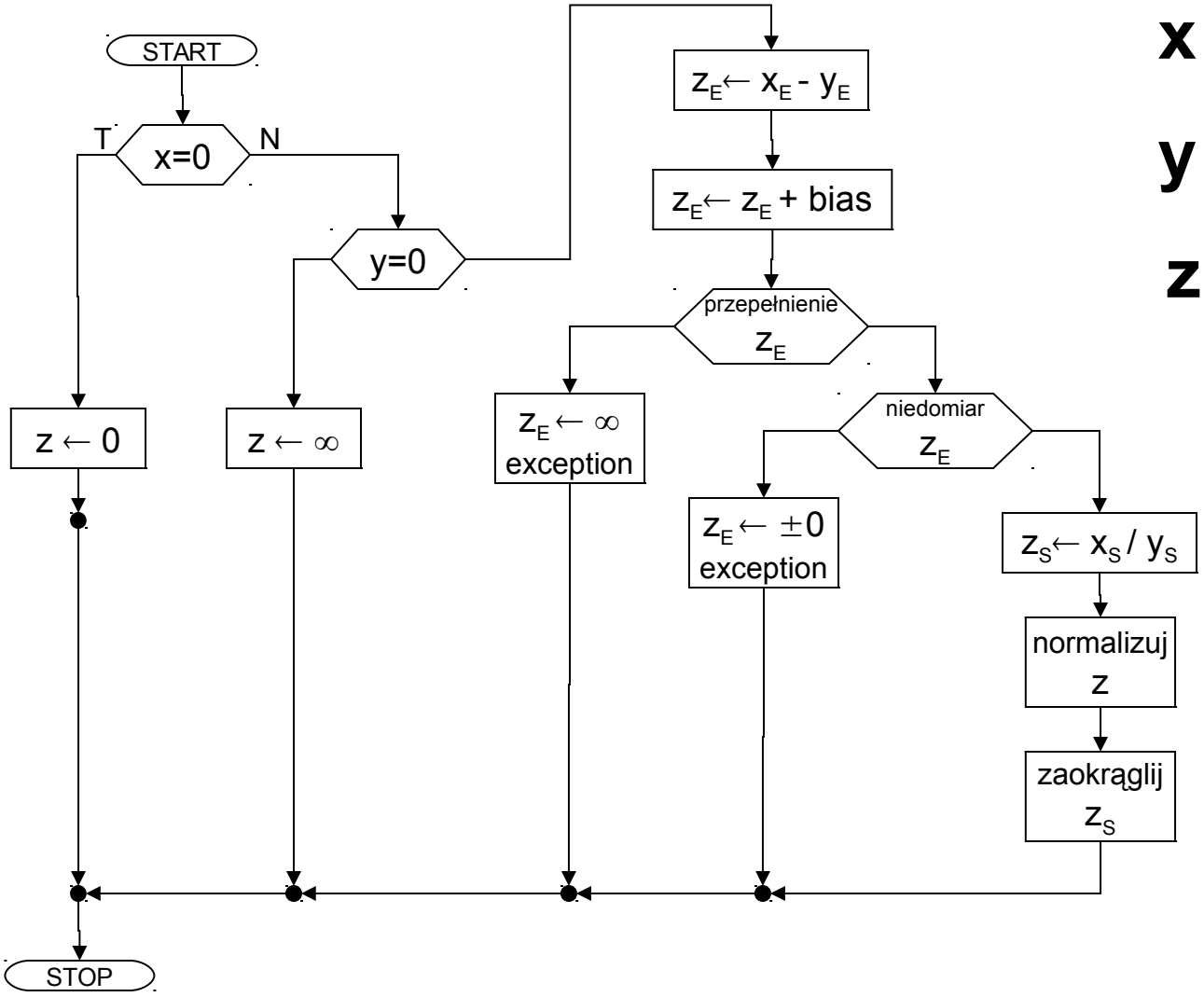


$$x = x_S \cdot 2^{x_E}$$

$$y = y_S \cdot 2^{y_E}$$

$$z = x_S \cdot y_S \cdot 2^{x_E + y_E}$$

Algorytm dzielenia



$$x = x_S \cdot 2^{x_E}$$

$$y = y_S \cdot 2^{y_E}$$

$$z = x_S / y_S \cdot 2^{x_E - y_E}$$