RC 5246

Technische Hogeschool Eindhoven
Onderafdeling der Wiskunde
Rekencentrum

RC-Informatie nr. 21
8 bladzijden
datum: 1 december 1969

ALGOL procedures voor het rekenen in dubbele lengte

1. Inleiding

In de meeste machines worden de reals uit een ALGOL programma voorgesteld in zg. floating point representatie. Voor een binaire machine betekent dit dat de waarde van een real steeds van de vorm

$$x = m \times 2^p \qquad (1)$$

is, met

$$m \text{ geheel}, \ |m| \leq 2^t - 1 \qquad \text{(de mantisse)}$$

$$p \text{ geheel}, \ -pmin \leq p \leq pmax \qquad \text{(de exponent)}.$$

Hierin zijn t, pmin, pmax bij de machine behorende positieve getallen; t is de mantisselengte, het aantal bits nodig voor de representatie van m. Voor de EL X8 is $t = 40$, $pmin = pmax = 2047 \ (= 2^{11} - 1)$.

De getallen van de vorm (1) noemen we machinegetallen.
Uiteraard is niet ieder reëel (of rationaal) getal een machinegetal. Wel is er bij ieder reëel getal x een machinegetal x1 dat een beste afronding van x tot een machinegetal is.
Deze x1 is gekarakteriseerd door

$$|x - x1| \leq |x - y| \quad \text{voor alle machinegetallen } y.$$

(Als x precies het gemiddelde van twee opeenvolgende machinegetallen is, dan is x1 hierdoor niet eenduidig bepaald, vandaar dat we spreken van <u>een</u> beste afronding.)

Voor het verschil tussen x1 en x (de afrondingsfout) geldt:
Als $2^{p+t-1} \leq |x| \leq 2^{p+t}$ met p geheel en $-pmin \leq p \leq pmax$, dan voldoet (voldoen) de beste afronding(en) x1 van x aan

$$|x - x1| \leq \tfrac{1}{2} \times 2^p$$

waaruit volgt dat altijd geldt:

$$|x - x1| \leq 2^{-t} \times |x| \ .$$

We laten de extreem grote getallen x (met $|x| > 2^{t+pmax}$) en de extreem dicht bij nul liggende getallen x (met $0 < |x| < 2^{t-pmin-1}$) verder buiten beschouwing. Dan is de relatieve afrondingsfout bij afronding van een reëel getal $\neq 0$ tot een machinegetal dus hoogstens 2^{-t} (voor de EL X8 dus $2^{-40} \sim .9 \times 10^{-12}$).

Het exacte resultaat van optelling, aftrekking, vermenigvuldiging of deling van twee machinegetallen is als regel niet een machinegetal.
Bij uitvoering van een statement als z := x+y moet dus als regel afgerond worden. We nemen aan dat de machine steeds correct afrondt (bij de EL X8 is dit zo). Dan is de relatieve fout bij optelling, etc., ook hoogstens 2^{-t}.

Voor de meeste toepassingen is een relatieve nauwkeurigheid in de elementaire operaties (+,-,×,/) van 2^{-t} ruimschoots voldoende. Het komt echter voor dat men een deel van een berekening met grotere relatieve nauwkeurigheid wenst uit te voeren.
Bijvoorbeeld is dit het geval als men een gewenst resultaat alleen kan verkrijgen als verschil van twee zeer veel grotere tussenresultaten.
Men kan dan voldoende nauwkeurigheid in het eindresultaat alleen verkrijgen indien de tussenresultaten met extra nauwkeurigheid berekend zijn.

Het is mogelijk de elementaire operaties met een relatieve nauwkeurigheid van ongeveer 2^{2t} uit te voeren, door een reëel getal te benaderen door een combinatie van twee machinegetallen, een zg. dubbellengtepaar.

We noemen een paar getallen x1 en x0 een dubbellengtepaar {x1,x0} indien:
a) x1 en x0 zijn machinegetallen,
b) voor alle machinegetallen y geldt

$$|(x1 + x0) - x1| \leq |(x1 + x0) - y| ,$$

d.w.z. x1 is een beste afronding van (x1 + x0) tot machinegetal.
De waarde van het dubbellengtepaar {x1,x0} is per definitie x1 + x0.
(N.b.: Er wordt niet geëist dat x1 en x0 hetzelfde teken hebben!)

Men kan bewijzen dat er bij ieder reëel getal x (dat niet extreem groot is en niet extreem dicht bij 0 ligt) een dubbellengtepaar {x1,x0} is zo, dat

$$|x - (x1 + x0)| \leq 2^{-2t-1} \times |x| .$$

Hieronder volgen procedures voor:

a) het optellen, vermenigvuldigen, delen van machinegetallen x en y met als resultaat een dubbellengtepaar {z1,z0}: DLA11, DLM11, DLD11;

b) het optellen, vermenigvuldigen, delen van een dubbellengtepaar {x1,x0} en een machinegetal y met als resultaat een dubbellengtepaar {z1,z0}: DLA21, DLM21, DLD21;

c) het delen van een machinegetal x door een dubbellengtepaar {y1,y0} met als resultaat een dubbellengtepaar {z1,z0}: DLD12;

d) het optellen, vermenigvuldigen, delen van dubbellengteparen {x1,x0} en {y1,y0} met als resultaat een dubbellengtepaar {z1,z0}: DLA22, DLM22, DLD22.

Bij de procedures ad a) wordt de maximaal bereikbare nauwkeurigheid gehaald (optelling en vermenigvuldiging zijn exact, bij deling is de relatieve fout hoogstens 2^{-2t-1}). Bij de procedures ad b), c) en d) is de relatieve fout nooit groter dan enkele malen 2^{-2t}.

De procedures zijn geheel in ALGOL geschreven. Zij functioneren correct in elke machine die bij optelling en vermenigvuldiging van machinegetallen steeds een beste afronding aflevert (bij een andere mantisselengte dan 40 moet de factor 1048577 in de procedures DLM en DLD vervangen worden door $2^s + 1$ met $s = ((t+1) \div 2)$.

De tijd nodig voor de uitvoering van een procedure aanroep ligt tussen 1 milliseconde (DLA11) en 2.5 milliseconde (DLD22).

2. Gebruiksaanwijzing

procedure DLA11(x,y,z1,z0);	{z1,z0} := x + y;
procedure DLM11(x,y,z1,z0);	{z1,z0} := x × y;
procedure DLD11(x,y,z1,z0);	{z1,z0} := x/y;
procedure DLA21(x1,x0,y,z1,z0);	{z1,z0} := {x1,x0} + y;
procedure DLM21(x1,x0,y,z1,z0);	{z1,z0} := {x1,x0} × y;
procedure DLD21(x1,x0,y,z1,z0);	{z1,z0} := {x1,x0} / y;
procedure DLD12(x,y1,y0,z1,z0);	{z1,z0} := x / {y1,y0};
procedure DLA22(x1,x0,y1,y0,z1,z0);	{z1,z0} := {x1,x0} + {y1,y0};
procedure DLM22(x1,x0,y1,y0,z1,z0);	{z1,z0} := {x1,x0} × {y1,y0};
procedure DLD22(x1,x0,y1,y0,z1,z0);	{z1,z0} := {x1,x0} / {y1,y0};

Alle formele parameters zijn real gespecificeerd.
Alle formele parameters behalve z1 en z0 worden by value aangeroepen.
Het is toegestaan dat de actuele parameter voor z1 of z0 dezelfde is als een van de overige actuele parameters.
Het is noodzakelijk dat de paren x1,x0 en y1,y0 dubbellengteparen zijn.
Het afgeleverde paar z1,z0 is een dubbellengtepaar.
Zonodig kan van een paar x1,x0 een dubbellengtepaar gemaakt worden (met waarde x1 + x0) door de aanroep: DLA11(x1,x0,x1,x0).
Een paar x1,x0 met x0 = 0 vormt altijd een dubbellengtepaar.

3. ALGOL-tekst

```
procedure DLA11(x, y, z1, z0);
   value x, y; real x, y, z0, z1;
   begin real z;
         z1:= z := x + y; z0:= if abs(x) > abs(y) then x - z + y else y - z + x
   end DLA11;

   procedure DLA21(x1, x0, y, z1, z0);
   value x1, x0, y; real x1, x0, y, z1, z0;
   begin real u1, u0, z;
         u1:= x1 + y;
         u0:= if abs(x1) > abs(y) then x1 - u1 + y + x0 else y - u1 + x1 + x0;
         z1:= z := u1 + u0; z0:= u1 - z + u0
   end DLA21;

   procedure DLA22(x1, x0, y1, y0, z1, z0);
   value x1, x0, y1, y0; real x1, x0, y1, y0, z1, z0;
   begin real z, u0, u1, v0, v1;
         u1 := x1 + y1;
         u0 := if abs(x1) > abs(y1) then x1 - u1 + y1 else y1 - u1 + x1;
         if u0 = 0 then
         begin u0 := x0 + y0; v1 := u1 + u0;
               v0 := u1 - v1 + u0 +
                     (if abs(x0) > abs(y0) then x0 - u0 + y0 else y0 - u0 + x0);
               z1 := z := v1 + v0; z0 := v1 - z + v0
         end else
         begin u0 := u0 + x0 + y0; z1 := z := u1 + u0; z0 := u1 - z + u0 end
   end DLA22;

   procedure DLM11(x, y, z1, z0);
   value x, y; real x, y, z1, z0;
   if x = 0 ∨ y = 0 then z1 := z0 := 0 else
   begin real x1, y1, z;
         z1 := z := x × y;
         x1 := 1048577 × x; x1 := x - x1 + x1; x := x - x1;
         y1 := 1048577 × y; y1 := y - y1 + y1; y := y - y1;
         z0 := x1 × y1 - z + x1 × y + x × y1 + x × y
   end DLM11;
```

```
procedure DLM21(x1, x0, y, z1, z0);
value x1, x0, y; real x1, x0, y, z1, z0;
if x1 = 0 ∨ y = 0 then z1 := z0 := 0 else
begin real u1, u0, x, z;
      u1 := x1 × y; u0 := x0 × y;
      x := 1048577 × x1; x := x1 - x + x; x1 := x1 - x;
      z := 1048577 × y; z := y - z + z; y := y - z;
      u0 := x × z - u1 + x × y + x1 × z + x1 × y + u0;
      z1 := z := u1 + u0; z0 := u1 - z + u0
end DLM21;

procedure DLM22(x1, x0, y1, y0, z1, z0);
value x1, x0, y1, y0; real x1, x0, y1, y0, z1, z0;
if x1 = 0 ∨ y1 = 0 then z1 := z0 := 0 else
begin real u1, u0, z;
      u1 := x1 × y1; u0 := x1 × y0 + x0 × y1;
      x0 := 1048577 × x1; x0 := x1 - x0 + x0; x1 := x1 - x0;
      y0 := 1048577 × y1; y0 := y1 - y0 + y0; y1 := y1 - y0;
      u0 := x0 × y0 - u1 + x0 × y1 + x1 × y0 + x1 × y1 + u0;
      z1 := z := u1 + u0; z0 := u1 - z + u0
end DLM22;

procedure DLD11(x, y, z1, z0);
value x, y; real x, y, z1, z0;
if y = 0 then begin z1 := x/y; z0 := 0 end else
if x = 0 then z1 := z0 := 0 else
begin real u1, u0, v1, v0, z;
      z1 := z := x/y;
      u1 := z × y;
      u0 := 1048577 × z; u0 := z - u0 + u0; z := z - u0;
      v1 := 1048577 × y; v1 := y - v1 + v1; v0 := y - v1;
      u0 := u0 × v1 - u1 + u0 × v0 + v1 × z + v0 × z;
      z0 := (x - u1 - u0)/y
end DLD11;
```

```
procedure DLD21(x1, x0, y, z1, z0);
value x1, x0, y; real x1, x0, y, z1, z0;
if y = 0 then begin z1 := x1/y; z0 := 0 end else
if x1 = 0 then z1 := z0 := 0 else
begin real u1, u0, v1, v0, z, u;
      u1 := z := x1/y;
      u := z × y;
      u0 := 1048577 × z; u0 := z - u0 + u0; z := z - u0;
      v1 := 1048577 × y; v1 := y - v1 + v1; v0 := y - v1;
      u0 := u0 × v1 - u + u0 × v0 + v1 × z + v0 × z;
      u0 := (x1 - u - u0 + x0)/y;
      z1 := z := u1 + u0; z0 := u1 - z + u0
end DLD21;

procedure DLD12(x, y1, y0, z1, z0);
value x, y1, y0; real x, y1, y0, z1, z0;
if y1 = 0 then begin z1 := x/y1; z0 := 0 end else
if x = 0 then z1 := z0 := 0 else
begin real u1, u0, v1, v0, z, u;
      u1 := z := x/y1;
      u := z × y1;
      u0 := 1048577 × z; u0 := z - u0 + u0; z := z - u0;
      v1 := 1048577 × y1; v1 := y1 - v1 + v1; v0 := y1 - v1;
      u0 := u0 × v1 - u + u0 × v0 + v1 × z + v0 × z;
      u0 := (x - u - u0 - y0 × u1)/y1;
      z1 := z := u1 + u0; z0 := u1 - z + u0
end DLD12;

procedure DLD22(x1, x0, y1, y0, z1, z0);
value x1, x0, y1, y0; real x1, x0, y1, y0, z1, z0;
if y1 = 0 then begin z1 := x1 / y1; z0 := 0 end else
if x1 = 0 then z1 := z0 := 0 else
begin real u1, u0, v1, v0, z, u;
      u1 := z := x1/y1;
      u := z × y1;
      u0 := 1048577 × z; u0 := z - u0 + u0; z := z - u0;
      v1 := 1048577 × y1; v1 := y1 - v1 + v1; v0 := y1 - v1;
      u0 := u0 × v1 - u + u0 × v0 + v1 × z + v0 × z;
      u0 := (x1 - u - u0 + x0 - y0 × u1)/y1;
      z1 := z := u1 + u0; z0 := u1 - z + u0
end DLD22;
```

4. Nauwkeurigheid

DLA11	$\{z1,z0\} := x + y$	exact
DLA21	$\{z1,z0\} := \{x1,x0\} + y$	$\|fout\| \leq 2 \times 2^{-2t} \times \|x1+x0+y\|$
→ DLA22	$\{z1,z0\} := \{x1,x0\} + \{y1,y0\}$	$\|fout\| \leq 3 \times 2^{-2t} \times \|x1+x0+y1+y0\|$
⟼ DLM11	$\{z1,z0\} := x \times y$	exact
DLM21	$\{z1,z0\} := \{x1,x0\} \times y$	$\|fout\| \leq 3 \times 2^{-2t} \times \|x1+x0\| \times \|y\|$
→ DLM22	$\{z1,z0\} := \{x1,x0\} \times \{y1,y0\}$	$\|fout\| \leq 7 \times 2^{-2t} \times \|x1+x0\| \times \|y1+y0\|$
DLD11	$\{z1,z0\} := x / y$	$\|fout\| \leq \frac{1}{2} \times 2^{-2t} \times \|x/y\|$
DLD21	$\{z1,z0\} := \{x1,x0\} / y$	$\|fout\| \leq 4 \times 2^{-2t} \times \|x1+x0\| / \|y\|$
DLD12	$\{z1,z0\} := x / \{y1,y0\}$	$\|fout\| \leq 7 \times 2^{-2t} \times \|x\| / \|y1+y0\|$
→ DLD22	$\{z1,z0\} := \{x1,x0\} / \{y1,y0\}$	$\|fout\| \leq 12 \times 2^{-2t} \times \|x1+x0\| / \|y1+y0\|$

Dekker

add 2 }
sub 2 }

mul 2

div 2

sqrt 2

$2^{2-2t}(|x1+x0| + |y1+y0|)$

9×2^{-2t}

12.1×2^{-2t}

10.2×2^{-2t}