DIGITAL ARITHMETIC

Miloš D. Ercegovac and Tomás Lang

Morgan Kaufmann Publishers, an imprint of Elsevier Science, ©2004

COMMENTS AND ERRATA

Updated: March 14, 2005

Chapter 1

page 36, line 14 and 16: replace $\tilde{w}[j]$ with $\tilde{w}[j+1]$

page 39, lines 8, 10, and -4: Replace q_{n-1-j} with q_{n-j}

page 39, in Algorithm NRD, move **endfor** before Step 4.

page 39, line 17: replace "fractions" with "integers"

page 45: Exercise 1.22: replace "right" with "left"

page 46, line -5: replace "Hennessay" with "Hennessy"

Chapter 2

page 64, line -2: replace T_{SRA} with T_{SCRA}

page 64, line -1: after "... buffers required" insert "for"

page 65, line 4: replace "(2.21)" with "(2.22)"

page 70, line 1 in Section 2.5.2: (2.30) should be (2.31)

Comment on delay calculation for Carry-lookahead adder (CLA): Because of the implementation of the CLG module shown in Figure 2.14, in delay expressions 2.43, 2.48, and 2.52 the term t_{clg} corresponds to the delay between module input c_0 and module output c_i, $i \neq 0$.

page 82, line -11: Change "The number of cells is the same as for the basic scheme." to "The number of cells is reduced by two compared to the basic scheme."

page 89: include the delay of buffers t_{buff} in the expression (2.72).

page 115: Exercise 2.2: replace "Table 2.4" with "Table 2.2"

Chapter 3

page 154, in Figure 3.13, add to the caption: (for $p = 3$)

page 155, in Figure 3.14, line 13: replace "d xxxx" with "d xxx" (only 3 x's)

page 156, line -11: replace $T_{[4:2]} < 2T_{[3:2]}$ with $t_{[4:2]} < 2t_{[3:2]}$

Chapter 4

page 183, line 1: insert "multiplying" between "for" and "magnitudes".

page 186, Figure 4.3: eliminate "Stage 3" in the top left corner.

page 193, line -9: replace "page 286" with "page 198"

page 196, Figure 4.9(b,c), column 4, row 4: replace x_l with x_1

page 216, expression (4.41): put parentheses around superscripts in the rightmost term.

Comment on Figure 4.26 (page 220): The figure is generic: it does not imply that for a 8-bit result $k = 6$. The references cited on pages 236-237 should be consulted for details about determining truncation precision.

page 227, line -7: replace with "For the design exercises use the circuit data from Table 2.4 and Figure 5.4. Note also that the delays in Figure 5.4 are given in t_{NAND2} units, whereas those in Table 2.4 are in nanoseconds. It might be best to report the results of the exercises in t_{NAND2} units.

page 229, in Exercise 4.8, replace "Exercise 4.7" with "Figure 4.7"

Chapter 5

page 264, line 2: replace "$ps + pc > 0$" with "$ps + pc < 0$"

page 264: in "Radix-4 division algorithm with the residual in carry-save form" there are some additional differences with respect to the radix-2 algorithm in Figure 5.6 which should be considered. Namely,

a) In the recurrence step, the number of iterations is $N = \lceil \frac{n+2+1}{2} \rceil$ (because of the initialization step and the guard bit) where n is the number of bits of the operands.

b) In the termination step, instead of the radix-2 expression

$$q = 2\,(CONVERT\,(Q\,[n+1]\,, q_{n+2} - 1))$$

use the corresponding radix-4 expression

$$q = 4\,(CONVERT\,(Q\,[N-1]\,, q_N - 1))$$

and instead of the expression

$$q = 2\left(CONVERT\left(Q\left[n+1\right], q_{n+2}\right)\right)$$

use

$$q = 4\left(CONVERT\left(Q\left[N-1\right], q_{N}\right)\right)$$

page 266, Figure 5.6: The fact that $\hat{y}$ is represented by three integer bits and one fractional bit, does not imply that its range is [-4, 3.5]. The range of $\hat{y}$ in the selection function is obtained from expression (5.102) and determined by expression (5.107) for radix-2 division with carry-save adder.

page 268, Figure 5.8: the output of the module "SZ; Convert" should be q.

page 270, Figure 5.9, line 14: the least significant bit of $4WC[2]$ should be 0, eliminate *

page 270, Figure 5.9, line 16: the least significant bit of $w[3]$ should be 0

page 278: line -4, replace "$0 \leq d \leq r^n - 1$" with "$0 < d \leq r^n - 1$"

page 278: In Table 5.8 footnote replace "Correction" by "Termination step"

page 278: In Section 5.4 there is ambiguity because of the use of r (radix) for two different purposes:

- On page 278 lines -5 and -4, the r refers to the radix in the representation of the operands. Usually, this radix will be 2. This also corresponds to the r in expression 5.46, 5.47, and 5.49.
- On page 279, line after expression 5.45, the $r = 2^k$ refers to the radix of the quotient-digit, as produced by the division algorithm. That, is for example in a radix-4 division algorithm, this radix would be 4.

To avoid this ambiguity, the caption of Figure 5.15 should say $n = 8$ bits, instead of $n = 4$ (radix-4 digits), since the operands are in radix 2.

page 279: Expression 5.44: replace the term "$... = 2^m \lfloor x/d^* \rfloor$" with

$$... = \lfloor 2^m \times (x/d^*) \rfloor$$

page 279: replace expression 5.48 with

$$N = \lceil (m+v)/k \rceil$$

That is, the $+1$ is incorrect and the k is missing. The argument for eliminating the $+1$ is based on the fact that the quotient obtained by a fractional division algorithm is $1/2 \leq q < 2$, as indicated in item 2 of the same page.

page 280: Example 5.1: replace the expression for N with

$$N = \lceil (m+v)/2 \rceil = 4$$

page 282: expression 5.53 should be $|w[j]| \leq \rho d$.

page 284: Figure 5.16(b): the label on X-axis should be d.

page 288: line 14: replace $\{r[w]\}_c$ with $\{rw[j]\}_c$

page 297: Figure 5.25, number the m_k constants beginning from $m_2(0)$ to $m_2(7)$.

page 311: Exercise 5.8: line -11: Remove sentence "Show all details.".

page 312: Exercise 5.15: line -9: Replace sentence "Draw the corresponding P-D diagrams (first quadrant only)." with "Draw the corresponding P-D diagrams (give the portion for $k = 6$, first quadrant only)."

page 313: Exercise 5.17: line 8: Expression $q_{j+1} = integer\,(rw\,[j] + 0.5)$ is valid if $w\,[j]$ is expressed in two's complement. When considering a sign and magnitude representation for the residuals, the expression has to be replaced by $q_{j+1} = round\,(rw\,[j])$.

page 313: Exercise 5.17: line 14: Replace "a fast radix-2 division algorithm." with "other low radix division algorithms.".

Chapter 6

page 352: line 15: replace sentence "$\max\left(L_k\left(I_i\right)\right)$ j is positive" with "$\max\left(L_k\left(I_i\right)\right)$ the term depending on j is positive".

page 358: Exercise 6.5, line 16: the expression for t_{cycle} is

$$t_{cycle} = t_{SEL_{SQRT}} + t_{buff} + t_{mux} + t_{HA} + t_{reg} = 4 + 1 + 1 + 1 + 2 = 9t_g$$

page 359: Exercise 6.8, line 6: replace sentence "Perform the integer division algorithm for radix 4 with residual in carry-save representation for $x = 53$ and $d = 9$." with "Perform the integer square root algorithm for radix 4 with residual in carry-save representation for $x = 53$."

Chapter 7

page 371, expression (7.11): replace $P[j]$ with $R[j]$

page 382, line 11: replace $R[j]$ with $S[j]$

page 388: Exercise 7.6: a and b are defined in Exercise 7.5.

Chapter 8

page 407, line 9: replace "are not representable in the floating-point system" with "do not correspond to real numbers"

page 420, footnote No. 14: replace "bised" with "biased"

page 427, line 9: replace "put" with "plus".

page 427, line 9, line -2: replace "complemented" with "negated" (two instances)

page 428: In Figure 8.8 the Exponent Update module should have also an input to update when the significand is shifted after the adder.

page 434, Paragraph 3, line 4: replace "significants" with "significands".

Chapter 9

pages 499 and 501, Figure 9.6 and 9.7: Replace "Shift-Reg WC" with "Reg WC"; replace "Shift-Reg WS" with "Reg WS"; replace "2w[j] with "w[j]"

page 510, line 6: eliminate one "the"

page 516, Table 9.4, row for $r = 2$: replace 6 with 3. Add to caption: The initial number of bits/operand is $log_2 r \times \delta$.

page 535: In Exercise 9.3 the initial conditions are $x[-1] = y[-1] = w[-1] = 0$. In the illustration of the input sequence, replace $x[j]$ and $y[j]$ with x_j and y_j, respectively.

Chapter 10

page 572: in the expression for the SEL digit selection function, replace the left-hand side with dk_{j+1}, and in the part with "-1", replace $\widehat{wk[j]}$ with $\widehat{vk[j]}$

Chapter 11

page 620: The sequence of scaling iterations is incorrect. A suitable sequence is (-1)(+2)(-5)(+8)(-10)(+15)(-17)(-19). Note that the scaling (-1) corresponds to a multiplication by 2^{-1}, so no scaling iteration is required.

The sequence of scalings plus repetitions is correct. However, it assumes that the CORDIC iterations begin at $j = 1$. This is acceptable because the repetitions make the convergence range as large as that without repetitions beginning at $j = 0$.

page 627: Table 11.4. For clarity and consistency the initial values x_i, y_i, and z_i should be denoted with x_{in}, y_{in}, and z_{in}, respectively.

page 628: row 7, last column of Table 11.5 should read: $z_R = 0.5\ln(4a)$

page 629, line 16: replace "hight" by "high"

page 629, last line: add "(See Exercise 11.4)"

page 630, replace lines -6 to -3 by: It has been shown that when m digits are used for the estimation of the sign, the distance between repetitions is $m-2$ iterations for the rotation mode and $m-5$ iterations for the vectoring mode.

More specifically, in iteration j the following digits are inspected:

- For rotation inspect digits with weights from $2^{-(j-1)}$ to $2^{-(j+m-2)}$.
- For vectoring inspect digits with weights from $2^{-(j-2)}$ to $2^{-(j+m-3)}$.

page 635: Exercise 11.1. Interpret "a precision of seven bits" as "perform the minimum number of iterations required to reduce the angle to 0 with the given data-path width". With respect to part c) we have not found a systematic solution.

page 636: Exercise 11.2. Perform the minimum number of iterations required to reduce the angle to 0 with the given datapath width. With respect to part c) we have not found a systematic solution.

page 636: Exercise 11.3. Perform the minimum number of iterations to reduce y to 0 with the given datapath width. With respect to part c) we have not found a systematic solution.

page 636: Exercise 11.4. The sequence α_i should be a decreasing sequence. That is, the relation between α_i and α_{i+1} should be

$$\alpha_{i+1} < \alpha_i \leq 2\alpha_{i+1}$$

page 637: Exercise 11.15 According to the Errata for page 630 replace "Use a selection function with an estimate of the sign with two digits..." by "... with four digits..."

page 637: Exercise 11.16 According to the Errata for page 630 use an estimate of seven digits.