COMPANY CONFIDENTIAL

PROJECT STRETCH

FILE MEMORANDUM # 35

SUBJECT: Practical Multiple Precision Division

BY: W. G. Bouricius

DATE: June 6, 1956

Mr. W. Wolensky in File Memorandum #36 discusses the theoretical possibilities of coding multiple precision division. Nothing in this memo disagrees with any of his conclusions, but it does point out some practical things regarding multiple precision division (plus implicitely also multiplication).

To program a code to find $\frac{1}{B_1 + B_2}$ we can use the following equations:

1. $$\frac{1}{B_1 + B_2} = \frac{1}{B_1}\left(\frac{1}{1 + \frac{B_2}{B_1}}\right) = \frac{1}{B_1}\left(1 - \frac{B_2}{B_1} + \frac{(B_2)^2}{B_1} - \dots\right)$$

$$\cong \frac{1}{B_1}\left(1 - \frac{B_2}{B_1}\right)$$

Let us assume the following:

a.) Fraction is n bits long.
b.) Exponent of B_2 is n less than exponent of B_1.

Then we can solve equation 1 as follows:

a.) $\frac{1}{B_1}$ can be divided twice, thus getting $\sim$2n bits accuracy.

b.) $\frac{B_2}{B_1}$ can be divided once, getting $\sim$n bits accuracy.

c.) Since $\frac{B_2}{B_1} \cong 2^{-n}$, $1 - \frac{B_2}{B_1}$ is $\sim$2n bits accurate.

d.) $\frac{1}{B_1}\left(1 - \frac{B_2}{B_1}\right)$ is then $\sim$ 2n bits accurate.

It would have required about twice as many operations to yield 2n bits guaranteed accuracy, and this is the major point of this memo.

Similarly with triple precision, Wolensky's equation $\frac{A_1 + A_2}{B_1 + B_2} + B_3$

$$= (A_1 + A_2)\,(Z - B_3 Z^2 + B_3 Z^3 - B_3 Z \quad \ldots)$$

can be more simply expressed as

$$\frac{A_1 + A_2}{B_1 + B_2 + B_3} = \left(\frac{A_1 + A_2}{B_1}\right)\left(1 - \frac{B_2 + B_3}{B_1} + \left(\frac{B_2 + B_3}{B_1}\right)^2 - \ldots\right)$$

As a practical matter,

$$2. \quad \frac{1}{B_1 + B_2 + B_3} \simeq \frac{1}{B_1}\left(1 - \frac{B_2 + B_3}{B_1} + \left(\frac{B_2}{B_1}\right)^2\right)$$

then we can solve equation 2 as follows:

a.) $\frac{1}{B_1}$ can be divided three times, thus getting $\sim$3n bits accuracy.

b.) $\frac{B_2 + B_3}{B_1}$ can be divided twice, thus getting $\sim$ 2n bits accuracy.

c.) $\left(\frac{B_2}{B_1}\right)$ was obtained as part of (b) above, so

$\left(\frac{B_2}{B_1}\right)^2$ can be obtained by one multiplication and one obtains $\sim$ 1n bits accuracy.

d.) Since $\frac{(B_2)^2}{B_1} = 2^{-2n}$, and $\frac{B_2 + B_3}{B_1} = 2^{-n}$, the sum

$1 - \frac{B_2 + B_3}{B_1} + \left(\frac{B_2}{B_1}\right)^2$ is approximately 3n bits accurate.

The largest error term ignored was $\frac{2B_2B_3}{B_1^2} = 2.2^{-3n}$ or less.

The product $\frac{1}{B_1}\left(1 - \frac{B_2 + B_3}{B_1} + \left(\frac{B_2}{B_1}\right)^2\right)$ is probably accurate to 3n - 3 or 3n - 4 bits.

Again, if one wished to guarantee 3n bit accuracy, one would have had to approximately double the program length.

WGB:ai
6/6/56