FURTHER REMARKS ON REDUCING TRUNCATION ERRORS

Recently Jack M. Wolfe [1] proposed the use of cascaded accumulators to evaluate a sum of the form $S = \sum_{i=1}^{N} y_i$ when N is large and all the y's are of roughly the same order of magnitude. His intention was to alleviate the accumulation of rounding or truncation errors which otherwise occurs when S is evaluated in the straightforward way illustrated by the following FORTRAN program.

```
1  S = 0·0
2  DO 4 I = 1, N
3  YI = ···
4  S = S + YI
5  ····
```

The rounding or truncation in statement 4 could contribute to a loss of almost $\log_{10} N$ significant decimals in S. This would be important in those cases where the values of YI computed in statement 3 were correct to nearly full machine precision; otherwise the uncertainty in the YI's would swamp any additional error introduced in statement 4.

Of course, the simplest and fastest way to prevent such figure-loss is to accumulate S to double-precision. For example, in a FORTRAN IV program it would suffice to precede statement 1 above by the TYPE statement DOUBLE PRECISION S · The convenient accessibility of double-precision in many FORTRAN and some ALGOL compilers indicates that double-precision will soon be universally acceptable as a substitute for ingenuity in the solution of numerical problems.

In the meantime, programmers without easy access to double-precision arithmetic may be able to simulate it in the program above by a method far simpler than Wolfe's, provided they are using one of the electronic computers which normalize floating-point sums before rounding or truncating them. Among such machines are, for example, the I.B.M. 704, 709, 7090, 7094, 7040, 7044 and 360 (short word arithmetic).

The trick to be described below does not work on machines such as the I.B.M. 650, 1620, Univac 1107 and the Control Data 3600 which round or truncate floating-point sums to single precision before normalizing them.

In the following program S2 is an estimate of the error caused when S = T was last rounded or truncated, and is used in statement 13 to compensate for that error. The parentheses in statement 23 must not be omitted; they cause the difference (S−T) to be evaluated first and hence, in most cases, without error because the difference is normalized before it is rounded or truncated.

```
 1  S = 0.0
    S2 = 0.0
 2  DO 4 I = 1, N
 3  YI = ···
13  S2 = S2 + YI
    T = S + S2
23  S2 = (S−T) + S2
 4  S = T
 5  ····
```

Until double-precision arithmetic was made a standard feature of the FORTRAN language, the author and his students used this trick on a 7090 in FORTRAN II programs to perform quadrature, solve differential equations and sum infinite series.

REFERENCE:

1. WOLFE, J. M. Reducing truncation errors by programming. *Comm. ACM 7* (June 1964), 355–356.

W. KAHAN
University of Toronto
Toronto, Ontario, Canada

RECEIVED JULY, 1964