N66-10423

Technical Memorandum No. 33-142

Study of the Accuracy of the Double-Precision Arithmetic Operations on the IBM 7094 Computer

C. L. Lawson

jpl

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

July 15, 1963

Technical Memorandum No. 33-142

Study of the Accuracy of the Double-Precision Arithmetic Operations on the IBM 7094 Computer

C. L. Lawson

W. R. Hoover

W. R. Hoover, Chief
Computer Applications and Data Systems Section

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

July 15, 1963

Prepared Under Contract No. NAS 7-100
National Aeronautics & Space Administration

CONTENTS

TABLES

TABLES (CONT'D)

ABSTRACT

The IBM 7094 operations DFMP, DFDP, DFAD, and DFSB and the FORTRAN II library subroutines (DFMP), (DFDP), (DFAD), and (DFSB) were tested by applying each to 32,000 pairs of random arguments uniformly distributed between 0.5 and 1.0.

The operations DFAD and DFSB performed true 54-bit floating-point arithmetic. The operation DFMP gave a correctly rounded 54-bit result in only 3.6% of the cases, and the error ranged from –6.00 to 0.00 in units of the last bit position of the product. The relative error ranged from $-0.750 * 2^{-50}$ ($= -0.666 * 10^{-15}$) to 0.

The corresponding figures for DFDP were 47.2%, –3.48 to 2.72, and $-0.45 * 2^{-51}$ ($= -0.20 * 10^{-15}$) to $0.62 * 2^{-51}$ ($= 0.28 * 10^{-15}$).

The subroutines were less accurate on the average than the corresponding hardware operations; however, there were cases such as multiplication by 1.0 (or by any other exact power of 2) in which the reverse was true.

I. INTRODUCTION

The purpose of this study was to investigate the accuracy of the double-precision (D.P.) hardware operations DFMP, DFDP, DFAD, and DFSB on the IBM 7094 computer. For comparison, some of the tests were run also on the FORTRAN II library subroutines (DFMP), (DFDP), (DFAD), and (DFSB). The test programs are identified as C2C, C3C, and C4C, and each required 5½ min of machine time.

Since, apart from the underflow and overflow, the error in multiplication and division depends only on the fractional parts and not on the exponent parts or the signs, it is sufficient to check these operations on arguments in the half open interval [0.5, 1.0).

Each operation was applied to the same set of 32,000 pairs of normalized double-precision arguments lying strictly between 0.5 and 1.0. These arguments were pseudorandom and uniformly distributed. Some additional effort was expended for the purpose of partially randomizing the bit pattern of each of these arguments. Although this set of arguments is not particularly appropriate for studying addition and subtraction, the same runs were made for these operations. The operation DFSB was exact in all cases, and the operation DFAD was exact half the time and too small by 0.5 in units of the last bit position of the sum the other half of the time. Since this is the best that could be expected from these operations, we saw no reason to test them further.

The subroutine (DFSB) was exact in all cases, but the subroutine (DFAD) erred by as much as 2.0 in the last-bit position.

It should be noted that there is no theoretical basis for asserting that the uniform distribution or any other distribution describes the manner in which the fractional parts of arguments occur in actual problems. The uniform distribution was used because it was simplest.[1]

[1]For a discussion suggesting that small fractional parts occur more frequently than large fractional parts, see Hamming, R. W., *Numerical Methods for Scientists and Engineers.* New York, McGraw-Hill, 1962. Pp.35–39.

II. THE DOUBLE-PRECISION RANDOM NUMBERS

Each Double-Precision (D.P.) random number was constructed from four single-word fixed-point random numbers. The fixed-point random numbers were generated by the following standard method:

$$X_1 = 5^{15} * 2^{-35}$$

$$X_i = X_{i-1} * 5^{15} \bmod 1 \qquad i = 2,3,\cdots$$

The numbers so generated are approximately uniformly distributed between 0 and 1.

The i^{th} D.P. random number was constructed as follows:

High Order Word:

Sign = +	
Bits 1 to 9	10 000 000 1
Bits 10 to 22	High Order 13 Bits of X_{4i}
Bits 23 to 35	High Order 13 Bits of X_{4i-1}

Low Order Word:

Sign = +	
Bits 1 to 8	01 100 101
Bits 9 to 22	High Order 14 Bits of X_{4i-2}
Bits 23 to 35	High Order 13 Bits of X_{4i-3}

The numbers constructed in this way are in normalized 7094 double-precision form and are approximately uniformly distributed between ½ and 1.

III. COMPUTATION OF ERRORS AND RELATIVE ERRORS

The arguments were converted to the three-word format appropriate to the Jet Propulsion Laboratory's 70-bit arithmetic subroutines PROQ, QUOQ, SUMQ, and DIFQ, and then operated upon by these subroutines to obtain the "true" results. The result of the D.P. operation being tested was converted to the 70-bit format and the difference (D.P. result minus "true" result) was computed using DIFQ. Then this difference was either scaled to units of the last bit position of the D.P. result (to obtain the quantity called "E" in the accompanying tables) or it was divided by the true result to obtain the relative error R $*2^{-54}$.

IV. CLASSIFICATION OF ERRORS

The product of two of the test arguments must lie between ¼ and 1, and thus the exponent part of the product is octal 177 or 200. When the exponent part is 177, it means that a normalization shift has occurred. Since it was expected that this shift would cause larger errors, the results were tabulated separately for the 177 and 200 cases.

If the arguments were truly uniformly distributed in (0.5,1.0), then the 177 case would be expected to occur about 38.6% of the time.

$$\int_{.5}^{1.} (.5/x - .5)\, dx \,/\, .25 = .3863$$

In our test, this occurred 38.5% of the time.

The quotient of two of the test arguments must lie between ½ and 2, and thus the exponent part of the result is 200 or 201. The results were tabulated separately for these two cases. The 200 result, expected 50% of the time, occurred 49.9% of the time.

V. DISCUSSION OF TABLES

Tables 1 through 9 contain the data obtained from the tests of DFMP and (DFMP). It should be noted that the errors are all negative, and that the spread is greater for the cases in which the exponent part of the product is 177, i.e., when a final normalizing shift occurred. The accuracy of DFMP would be improved, on the average, if the magnitude of the result were always increased by adding 1.0 in the last bit position. This fact can sometimes be exploited in coding.

It would be desirable to know exactly the least upper bounds on the errors of the various operations. In Table 10, examples are given that exhibit the largest errors which we were able to produce in DFMP and (DFMP).

Tables 11 through 19 present data on the tests of DFDP and (DFDP). Here, the errors are nearly symmetrically distributed about zero, and a high percentage of the errors are near zero.

In Tables 20 and 21, some useful cumulative percentages are given that were computed from the previous tables. For instance, if one interprets 16 decimal place accuracy to mean that the relative error is below $0.5 * 10^{-16}$, then the percentage of cases in which DFMP and DFDP deliver 16 decimal place accuracy is about 8.3% and 63.0%, respectively, (Table 20); whereas 15 decimal place accuracy can almost be guaranteed.

Tables 22, 23, and 24 give data on DFAD and (DFAD). No statistics are given with these tables as it is believed that the set of arguments used do not constitute an appropriate set for estimating the general behavior of addition or subtraction.

Table 1. Classification of relative error in the 7094 operation DFMP

R[a] Relative error in units of 2^{-54}	Number of occurrences			Percent
	Exponent part of product		Total	
	177_8	200_8		
−9.5 to −9.0	2	0	2	0.006
−9.0 to −8.5	3	0	3	0.009
−8.5 to −7.5	18	0	18	0.06
−7.5 to −7.0	33	0	33	0.1
−7.0 to −6.5	81	0	81	0.3
−6.5 to −6.0	137	0	137	0.4
−6.0 to −5.5	238	0	238	0.7
−5.5 to −5.0	459	4	463	1.4
−5.0 to −4.5	734	35	769	2.4
−4.5 to −4.0	1091	164	1255	3.9
−4.0 to −3.5	1583	556	2139	6.7
−3.5 to −3.0	1918	1346	3264	10.2
−3.0 to −2.5	1936	2679	4615	14.4
−2.5 to −2.0	1700	4156	5856	18.3
−2.0 to −1.5	1266	4738	6004	18.8
−1.5 to −1.0	707	3738	4445	13.9
−1.0 to −0.5	348	1903	2251	7.0
−0.5 to 0.0	59	368	427	1.3
TOTAL	12,313	19,687	32,000	
Mean R	−3.12	−1.96	−2.41	
Std. Dev. R	1.27	0.80	1.16	
Mean $\|R\|$	3.12	1.96	2.41	
Std. Dev. $\|R\|$	1.27	0.80	1.16	
Median $\|R\|$ interval	3.0 to 3.5	1.5 to 2.0	2.0 to 2.5	

[a] $R = 2^{54} \cdot (Z_{7094} - Z_{TRUE}) / Z_{TRUE}$

Table 2. Largest relative error encountered in DFMP when exponent part of product was 177

Y1	200	403	235	366	145	734	154	367[a]
Y2	200	435	757	120	145	756	321	060
Computed product	177	441	522	701	144	327	502	324
Rounded true product	177	441	522	701	144	327	502	331
Relative error	$-9.27 \cdot 2^{-54} = -0.514 \cdot 10^{-15}$							

[a] Rows 1 to 4 are octal; row 5 is decimal

Table 3. Largest relative error encountered in DFMP when exponent part of product was 200

Y1	200	512	170	434	145	752	375	676
Y2	200	622	022	227	145	776	202	245
Computed product	200	403	236	352	145	005	630	600
Rounded true product	200	403	236	352	145	005	630	603
Relative error	$-5.22 \cdot 2^{-54} = -0.289 \cdot 10^{-15}$							

Table 4. Classification of errors in the 7094 operation DFMP

E[a]	Number of occurrences: Exponent part of product 177_8	200_8	Total	Percent
−6.0 to −5.5	1	0	1	0.003
−5.5 to −5.0	32	0	32	0.1
−5.0 to −4.5	167	0	167	0.5
−4.5 to −4.0	505	0	505	1.6
−4.0 to −3.5	1188	0	1188	3.7
−3.5 to −3.0	1826	0	1826	5.7
−3.0 to −2.5	2421	46	2467	7.7
−2.5 to −2.0	2475	1042	3517	11.0
−2.0 to −1.5	1861	4831	6692	20.9
−1.5 to −1.0	1233	7777	9010	28.2
−1.0 to −0.5	509	4928	5437	17.0
−0.5 to 0.0	95	1063	1158	3.6
TOTAL	12,313	19,687	32,000	
Mean E	−2.51	−1.25	−1.74	
Std. dev. E	0.93	0.46	0.92	
Mean |E|	2.51	1.25	1.74	
Std. dev. |E|	0.93	0.46	0.92	
Median |E| interval	2.0 to 2.5	1.0 to 1.5	1.5 to 2.0	

[a]E = Computed product minus true product in units of the last bit of the computed product.

Table 5. Largest error encountered in DFMP when exponent part of product was 177

Y1	200	442	643	413	145	762	467	550
Y2	200	466	610	631	145	733	432	365
Computed product	177	541	022	174	144	450	466	302
Rounded true product	177	541	022	174	144	450	466	310
Error	−5.51							

Table 6. Largest error encountered in DFMP when exponent part of product was 200

Y1	200	775	762	663	145	755	775	432
Y2	200	541	647	014	145	730	361	350
Computed product	200	540	332	222	145	454	040	015
Rounded true product	200	540	332	222	145	454	040	020
Error	−2.78							

Table 7. Classification of errors in the FORTRAN library subroutine (DFMP)

E[a]	Number of occurrences			Percent
	Exponent part of product		Total	
	177_8	200_8		
−9.5 to −9.0	2	0	2	0.006
−9.0 to −7.5	39	0	39	0.1
−7.5 to −7.0	41	0	41	0.1
−7.0 to −6.5	99	4	103	0.3
−6.5 to −6.0	218	30	248	0.8
−6.0 to −5.5	374	74	448	1.4
−5.5 to −5.0	468	124	592	1.9
−5.0 to −4.5	569	173	742	2.3
−4.5 to −4.0	626	272	898	2.8
−4.0 to −3.5	854	600	1454	4.5
−3.5 to −3.0	1082	1202	2284	7.1
−3.0 to −2.5	1589	2273	3862	12.1
2.5 to −2.0	1812	3170	4982	15.6
−2.0 to −1.5	1696	3839	5535	17.3
−1.5 to −1.0	1698	4382	6080	19.0
−1.0 to −0.5	939	2842	3781	11.8
−0.5 to 0.0	207	702	909	2.8
TOTAL	12,313	19,687	32,000	
Mean E	−2.73	−1.90	−2.22	
Std. dev. E	1.51	0.99	1.29	
Mean |E|	2.73	1.90	2.22	
Std. dev. |E|	1.51	0.99	1.29	
Median |E| interval	2.0 to 2.5	1.0 to 1.5	1.5 to 2.0	

[a]E = Computed product minus true product in units of the last bit of the computed product.

Table 8. Largest error encountered in the subroutine (DFMP) when exponent part of product was 177

Y1	200	632	340	463	145	732	225	437
Y2	200	421	736	206	145	763	032	501
Computed product	177	667	142	234	144	114	332	570
Rounded true product	177	667	142	234	144	114	332	601
Error	−9.26							

Table 9. Largest error encountered in the subroutine (DFMP) when exponent part of product was 200

Y1	200	512	170	434	145	752	375	676
Y2	200	622	022	227	145	776	202	245
Computed product	200	403	236	352	145	005	630	574
Rounded true product	200	403	236	352	145	005	630	603
Error	−6.64							

Table 10. Examples of large errors in DFMP[a]

Exponent 177								
Y	200	400	000	001	145	777	777	776
Y^2, True	177	400	000	004	144	000	000	003
	111	777	777	760	056	000	000	010
Y^2, 7094 DFMP	177	400	000	003	144	777	777	776
Y^2, Subroutine (DFMP)	Same as 7094 DFMP							
Error	$-5.999\ 999\ 94$							
Relative error	$-11.999\ 999\ 17 * 2^{-54} = -0.666 \bullet 10^{-15}$							
Exponent 200								
Y	200	600	000	001	145	777	777	774
Y^2, True	200	440	000	002	145	777	777	775
	112	777	777	760	057	000	000	020
Y^2, 7094 DFMP	200	440	000	002	145	777	777	773
Error	$-2.999\ 999\ 94$							
Relative error	$-5.33 * 2^{-54} = -0.296 * 10^{-15}$							
Y^2, Subroutine (DFMP)	200	440	000	002	145	777	777	772
Error	$-3.999\ 999\ 94$							
Relative error	$-7.11 \bullet 2^{-54} = -0.395 * 10^{-15}$							

[a]These examples were constructed to exhibit errors in the 7094 operation DFMP that are close to the unattainable upper bounds. These bounds are $|E| \leq 6$, $|R| \leq 12 \bullet 2^{-54}$ for exponent 177 and $|E| \leq 3$, $|R| \leq 6 \bullet 2^{-54}$ for exponent 200.

Table 11. Classification of relative error in the 7094 operation DFDP

R^a Relative error in units of 2^{-54}	Number of occurrences			Percent
	Exponent part of quotient		Total	
	200_8	201_8		
−4.0 to −3.5	0	4	4	0.01
−3.5 to −3.0	0	16	16	0.05
−2.5 to −2.0	1	296	297	0.9
−2.0 to −1.5	6	578	584	1.8
−3.0 to −2.5	0	82	82	0.3
−1.5 to −1.0	135	1174	1309	4.1
−1.0 to −0.5	955	1820	2775	8.7
−0.5 to 0.0	2333	2306	4639	14.5
0.0 to 0.5	3805	2618	6423	20.1
0.5 to 1.0	3895	2406	6301	19.7
1.0 to 1.5	2821	1971	4792	15.0
1.5 to 2.0	1396	1334	2730	8.5
2.0 to 2.5	495	765	1260	3.9
2.5 to 3.0	122	400	522	1.6
3.0 to 3.5	9	179	188	0.6
3.5 to 4.0	0	50	50	0.2
4.0 to 4.5	0	25	25	0.08
4.5 to 5.0	0	3	3	0.009
TOTAL	15,973	16,027	32,000	
Mean R	0.62	0.35	0.48	
Std. dev. R	0.76	1.21	1.02	
Mean |R|	0.79	1.01	0.90	
Std. dev. |R|	0.58	0.76	0.68	
Median |R| interval	0.5 to 1.0	0.5 to 1.0	0.5 to 1.0	

[a] $R = 2^{54} * (Z_{7094} - Z_{TRUE}) / Z_{TRUE}$

Table 12. Largest relative error encountered in DFDP when exponent part of quotient was 200

Y1	200	432	122	261	145	644	055	420
Y2	200	754	555	051	145	316	044	207
Computed quotient	200	445	151	311	145	670	753	014
Rounded true quotient	200	445	151	311	145	670	753	012
Relative error	$3.31 * 2^{-51} = 0.184 * 10^{-15}$							

Table 13. Largest relative error encountered in DFDP when exponent part of quotient was 201

Y1	200	427	531	210	145	717	563	334
Y2	200	400	426	171	145	356	536	536
Computed quotient	201	427	051	742	146	741	201	207
Rounded true quotient	201	427	051	742	146	741	201	204
Relative error	$4.99 * 2^{-51} = 0.277 * 10^{-15}$							

Table 14. Classification of errors in the 7094 operation DFDP

E[a]	Number of occurrences: Exponent part of quotient 200_8	201_8	Total	Percent
−3.5 to −3.0		1	1	0.003
−3.0 to −2.5		8	8	0.03
−2.5 to −2.0		15	15	0.05
−2.0 to −1.5	5	134	139	0.4
−1.5 to −1.0	12	718	730	2.3
−1.0 to −0.5	690	1944	2634	8.2
−0.5 to 0.0	2723	3456	6179	19.3
0.0 to 0.5	4847	4068	8915	27.9
0.5 to 1.0	4445	3247	7692	24.0
1.0 to 1.5	2459	1700	4159	13.0
1.5 to 2.0	724	604	1328	4.2
2.0 to 2.5	67	123	190	0.6
2.5 to 3.0	1	9	10	0.03
TOTAL	15,973	16,027	32,000	
Mean E	0.49	0.22	0.35	
Std. dev. E	0.60	0.76	0.70	
Mean |E|	0.62	0.63	0.63	
Std. dev. |E|	0.46	0.47	0.46	
Median |E| interval	0.5 to 1.0	0.5 to 1.0	0.5 to 1.0	

[a]E = Computed quotient minus true quotient in units of the last bit of the computed quotient

Table 15. Largest error encountered in DFDP when exponent part of quotient was 200

Y1	200	443	535	714	145	321	505	755
Y2	200	456	211	136	145	263	144	417
Computed quotient	200	756	044	366	145	661	751	440
Rounded true quotient	200	756	044	366	145	661	751	435
Error	2.51							

Table 16. Largest error encountered in DFDP when exponent part of quotient was 201

Y1	200	775	255	627	145	204	213	241
Y2	200	412	636	176	145	755	730	011
Computed quotient	201	750	550	713	146	232	724	252
Rounded true quotient	201	750	550	713	146	232	724	255
Error	−3.48							

Table 17. Classification of errors in the FORTRAN library subroutine (DFDP)

E[a]	Number of occurrences			Percent
	Exponent part of quotient		Total	
	200_8	201_8		
−7.5 to −7.0	1	0	1	0.003
−7.0 to −6.5	9	0	9	0.03
−6.5 to −6.0	8	0	8	0.03
−6.0 to −5.5	13	0	13	0.04
−5.5 to −5.0	17	0	17	0.05
−5.0 to −4.5	18	5	23	0.07
−4.5 to −4.0	44	11	55	0.2
−4.0 to −3.5	97	49	146	0.5
−3.5 to −3.0	196	109	305	1.0
−3.0 to −2.5	404	275	679	2.1
−2.5 to −2.0	742	534	1276	4.0
−2.0 to −1.5	1062	982	2044	6.4
−1.5 to −1.0	1416	1938	3354	10.5
−1.0 to −0.5	2046	3087	5133	16.0
−0.5 to 0.0	2957	3561	6518	20.4
0.0 to 0.5	3293	2802	6095	19.0
0.5 to 1.0	2102	1570	3672	11.5
1.0 to 1.5	938	665	1603	5.0
1.5 to 2.0	333	280	613	1.9
2.0 to 2.5	152	97	249	0.8
2.5 to 3.0	78	45	123	0.4
3.0 to 3.5	33	16	49	0.2
3.5 to 4.0	10	1	11	0.03
4.0 to 4.5	4	0	4	0.01
TOTAL	15,973	16,027	32,000	
Mean E	−0.34	−0.39	−0.36	
Std. dev. E	1.21	1.00	1.11	
Mean |E|	0.93	0.83	0.88	
Std. dev. |E|	0.85	0.69	0.78	
Median |E| interval	0.5 to 1.0	0.5 to 1.0	0.5 to 1.0	

[a]E = Computed product minus true quotient in units of the last bit of the computed quotient

Table 18. Largest error encountered in the subroutine (DFDP) when exponent part of quotient was 200

Y1	200	513	525	221	145	547	614	756
Y2	200	527	244	727	145	204	606	337
Computed quotient	200	756	474	314	145	511	600	070
Rounded true quotient	200	756	474	314	145	511	600	077
Error	−7.32							

Table 19. Largest error encountered in the subroutine (DFDP) when exponent part of quotient was 201

Y1	200	625	756	157	145	767	347	545
Y2	200	534	424	421	145	046	412	421
Computed quotient	201	452	133	236	146	404	514	172
Rounded true quotient	201	452	133	236	146	404	514	177
Error	−4.86							

Table 20. Percentage of cases in which the magnitude of the relative error in the product or quotient is less than .5 * 2^{-n}

n	$.5 * 2^{-n}$	DFMP	DFDP
54	$0.278 \cdot 10^{-16}$	1.3%	34.6%
53	$0.555 \cdot 10^{-16}$	8.3%	63.0%
52	$0.111 \cdot 10^{-15}$	41.0%	92.4%
51	$0.222 \cdot 10^{-15}$	90.6%	99.9%
50	$0.444 \cdot 10^{-15}$	99.8%	100.0%
49	$0.888 \cdot 10^{-15}$	100.0%	100.0%

Table 21. Percentage of cases in which the fractional part of the product or quotient contains at least n correct bits[a]

n	Equivalent number of decimal digits	DFMP	DFDP	(DFMP) Subroutine	(DFDP) Subroutine
54	16.26	3.6%	47.2%	2.8%	39.4%
53	15.95	20.6%	79.4%	14.6%	66.9%
52	15.65	69.7%	99.3%	50.9%	90.7%
51	15.35	97.8%	100.0%	90.2%	99.7%
50	15.05	100.0%	100.0%	99.9%	100.0%
49	14.75	100.0%	100.0%	100.0%	100.0%

[a]The magnitude of the error does not exceed 1 in the $(n + 1)^{ST}$ bit position

Table 22. Classification of errors in the 7094 operation DFAD

E[a]	Number of occurrences Exponent = 201_8 in all cases	Percent
−0.5	15791	49.3
0.0	16209	50.7
TOTAL	32,000	

[a]E = Computed sum minus true sum in units of the last bit of the computed sum.

Table 23. Classification of errors in the FORTRAN library subroutine (DFAD)

E[a]	Number of occurrences Exponent = 201_8 in all cases	Percent
−2.0	1019	3.2
−1.5	2006	6.3
−1.0	8102	25.3
−0.5	13785	43.1
0.0	7088	22.2
TOTAL	32,000	

[a]E = Computed sum minus true sum in units of the last bit of the computed sum.

Table 24. Largest error encountered in the subroutine (DFAD)

Y1	200 625 461 043	145 266 372 407
Y2	200 665 336 416	145 543 613 545
Computed sum	201 645 407 631	146 015 103 064
True sum	201 645 407 631	146 015 103 066
Error	−2.00	

VI. OTHER OBSERVATIONS

A. Quotient Not Normalized

Some exceptions were found to the statement that "The quotient is in normal form if both the dividend and divisor are in normal form."[2] A specific case in which the quotient produced was not normalized can be illustrated by the result obtained when Z was computed as X/Y:

X =	201 400 000 000	146 000 000 000
Y =	200 400 000 000	145 400 000 007
Z =	202 377 777 777	147 377 777 772

It seems reasonable to infer that this lack of normalization will occur whenever the high-order words of X and Y have the same fractional part, and the fractional part of the low-order word of Y exceeds the fractional part of the low-order word of X.

[2]*IBM 7094 Data Processing System Reference Manual,* No. A22-6703, Copyright 1962 by the International Business Machines Corporation, New York, p. 40.

[3]Hardware modifications made by IBM engineers in June 1963 have corrected this normalization problem.

It appears as though the DFDP operation should end with a test of normalization as the DFMP operation does. Unless this is corrected by hardware changes, it will be necessary to take into account the possibility of unnormalized data in output conversion routines, square root, logarithm, and other programs which have in the past taken advantage of the assumption that all arguments would be normalized.[3]

B. Multiplication: the 177 Case

It has been noted previously that about 38.6% of the time the fractional part of a product will be small enough to require a final normalizing shift during the DFMP operation. In such cases, the final bit position of the final result will necessarily be zero. It should be noted that multiplication by 1.0 or any other exact power of 2.0 will always produce this situation. Thus, for example, in multiplying by 1.0, using DFMP, the last bit of the result will always be a zero and thus will be wrong about half the time. This is one case in which (DFMP) excells DFMP as (DFMP) gives a correct product in this case.