

An 8-Bit Floating Point Representation

©2005 Dr. William T. Verts

In order to better understand the IEEE 754 floating point format, we use a simple example where we can exhaustively examine every possible bit pattern. An 8-bit format, although too small to be seriously practical, is both large enough to be instructive and small enough to be examined in its entirety. In this 8-bit format, one bit is reserved for the sign as usual, three bits are used for the biased exponent, and the remaining four bits are used for the mantissa. This format is shown below:



With 3 bits available, biased exponent values range between 0 (denormalized numbers, including zero) and 7 (infinity and NaN). For N bits of exponent the bias value is computed by the expression $2^{(N-1)}-1$, so in this case the bias is $2^{(3-1)}-1$, or 3. The range of legal exponents (those representing neither denormalized or infinite numbers) is then between 2^{-2} and 2^{+3} . For four bits of mantissa, the significant digits of any normalized non-zero binary number will range between 1.0000 and 1.1111, but we drop the always-present leading 1 bit in order to store five bits of precision into four bits of storage. With five bits of precision, the equivalent number of decimal digits is computed by the expression $\log_{10}(2^5)$, or approximately 1.505 digits.

The largest legal value in this format is 1.1111×2^3 , which is 1111.1 in binary or 15.5 in decimal. At the other end of the scale, the smallest normalized value is 1.0000×2^{-2} , which is 0.01 in binary or 0.25 in decimal. With denormalized numbers, we can pack more values between zero and the smallest legal normalized value, allowing for a gradual underflow to zero.

Notice that the difference between two adjacent entries in the list increases as the values get further and further away from zero. This is characteristic of all floating point systems: representable values are most densely packed close to zero, and spread out as they get further away. The information density is *not* uniform over the range of legal values.

Ignoring the sign, the integer values of the bits approximate the logarithm of the equivalent floating-point number. With the biased exponent representation, it is possible to compare two floating-point values of the same sign for relative magnitude by comparing their values as if they were integers. This is a practical concern for real assembly language programs since integer operations are typically much faster than floating-point operations.

N	Floating	Scientific	Binary	Decimal	Note
0	0-000-0000	+0.0000×2 ⁻²	+0.0	+0.0	Denormalized, +Zero
1	0-000-0001	+0.0001×2 ⁻²	+0.000001	+0.015625	Denormalized
2	0-000-0010	+0.0010×2 ⁻²	+0.00001	+0.03125	Denormalized
3	0-000-0011	+0.0011×2 ⁻²	+0.000011	+0.046875	Denormalized
4	0-000-0100	+0.0100×2 ⁻²	+0.0001	+0.0625	Denormalized
5	0-000-0101	+0.0101×2 ⁻²	+0.000101	+0.078125	Denormalized
6	0-000-0110	+0.0110×2 ⁻²	+0.00011	+0.09375	Denormalized
7	0-000-0111	+0.0111×2 ⁻²	+0.000111	+0.109375	Denormalized
8	0-000-1000	+0.1000×2 ⁻²	+0.001	+0.125	Denormalized
9	0-000-1001	+0.1001×2 ⁻²	+0.001001	+0.140625	Denormalized
10	0-000-1010	+0.1010×2 ⁻²	+0.00101	+0.15625	Denormalized
11	0-000-1011	+0.1011×2 ⁻²	+0.001011	+0.171875	Denormalized
12	0-000-1100	+0.1100×2 ⁻²	+0.0011	+0.1875	Denormalized
13	0-000-1101	+0.1101×2 ⁻²	+0.001101	+0.203125	Denormalized
14	0-000-1110	+0.1110×2 ⁻²	+0.00111	+0.21875	Denormalized
15	0-000-1111	+0.1111×2 ⁻²	+0.001111	+0.234375	Denormalized
16	0-001-0000	+1.0000×2 ⁻²	+0.01	+0.25	
17	0-001-0001	+1.0001×2 ⁻²	+0.010001	+0.265625	
18	0-001-0010	+1.0010×2 ⁻²	+0.01001	+0.28125	
19	0-001-0011	+1.0011×2 ⁻²	+0.010011	+0.296875	
20	0-001-0100	+1.0100×2 ⁻²	+0.0101	+0.3125	
21	0-001-0101	+1.0101×2 ⁻²	+0.010101	+0.328125	
22	0-001-0110	+1.0110×2 ⁻²	+0.01011	+0.34375	
23	0-001-0111	+1.0111×2 ⁻²	+0.010111	+0.359375	
24	0-001-1000	+1.1000×2 ⁻²	+0.011	+0.375	
25	0-001-1001	+1.1001×2 ⁻²	+0.011001	+0.390625	
26	0-001-1010	+1.1010×2 ⁻²	+0.01101	+0.40625	
27	0-001-1011	+1.1011×2 ⁻²	+0.011011	+0.421875	
28	0-001-1100	+1.1100×2 ⁻²	+0.0111	+0.4375	
29	0-001-1101	+1.1101×2 ⁻²	+0.011101	+0.453125	
30	0-001-1110	+1.1110×2 ⁻²	+0.01111	+0.46875	
31	0-001-1111	+1.1111×2 ⁻²	+0.011111	+0.484375	
32	0-010-0000	+1.0000×2 ⁻¹	+0.1	+0.5	
33	0-010-0001	+1.0001×2 ⁻¹	+0.10001	+0.53125	
34	0-010-0010	+1.0010×2 ⁻¹	+0.1001	+0.5625	
35	0-010-0011	+1.0011×2 ⁻¹	+0.10011	+0.59375	
36	0-010-0100	+1.0100×2 ⁻¹	+0.101	+0.625	
37	0-010-0101	+1.0101×2 ⁻¹	+0.10101	+0.65625	
38	0-010-0110	+1.0110×2 ⁻¹	+0.1011	+0.6875	
39	0-010-0111	+1.0111×2 ⁻¹	+0.10111	+0.71875	
40	0-010-1000	+1.1000×2 ⁻¹	+0.11	+0.75	
41	0-010-1001	+1.1001×2 ⁻¹	+0.11001	+0.78125	
42	0-010-1010	+1.1010×2 ⁻¹	+0.1101	+0.8125	
43	0-010-1011	+1.1011×2 ⁻¹	+0.11011	+0.84375	
44	0-010-1100	+1.1100×2 ⁻¹	+0.111	+0.875	
45	0-010-1101	+1.1101×2 ⁻¹	+0.11101	+0.90625	
46	0-010-1110	+1.1110×2 ⁻¹	+0.1111	+0.9375	
47	0-010-1111	+1.1111×2 ⁻¹	+0.11111	+0.96875	
48	0-011-0000	+1.0000×2 ⁰	+1.0	+1.0	
49	0-011-0001	+1.0001×2 ⁰	+1.0001	+1.0625	
50	0-011-0010	+1.0010×2 ⁰	+1.001	+1.125	
51	0-011-0011	+1.0011×2 ⁰	+1.0011	+1.1875	
52	0-011-0100	+1.0100×2 ⁰	+1.01	+1.25	
53	0-011-0101	+1.0101×2 ⁰	+1.0101	+1.3125	
54	0-011-0110	+1.0110×2 ⁰	+1.011	+1.375	
55	0-011-0111	+1.0111×2 ⁰	+1.0111	+1.4375	

56	0-011-1000	$+1.1000 \times 2^0$	+1.1	+1.5
57	0-011-1001	$+1.1001 \times 2^0$	+1.1001	+1.5625
58	0-011-1010	$+1.1010 \times 2^0$	+1.101	+1.625
59	0-011-1011	$+1.1011 \times 2^0$	+1.1011	+1.6875
60	0-011-1100	$+1.1100 \times 2^0$	+1.11	+1.75
61	0-011-1101	$+1.1101 \times 2^0$	+1.1101	+1.8125
62	0-011-1110	$+1.1110 \times 2^0$	+1.111	+1.875
63	0-011-1111	$+1.1111 \times 2^0$	+1.1111	+1.9375
64	0-100-0000	$+1.0000 \times 2^1$	+10.0	+2.0
65	0-100-0001	$+1.0001 \times 2^1$	+10.001	+2.125
66	0-100-0010	$+1.0010 \times 2^1$	+10.01	+2.25
67	0-100-0011	$+1.0011 \times 2^1$	+10.011	+2.375
68	0-100-0100	$+1.0100 \times 2^1$	+10.1	+2.5
69	0-100-0101	$+1.0101 \times 2^1$	+10.101	+2.625
70	0-100-0110	$+1.0110 \times 2^1$	+10.11	+2.75
71	0-100-0111	$+1.0111 \times 2^1$	+10.111	+2.875
72	0-100-1000	$+1.1000 \times 2^1$	+11.0	+3.0
73	0-100-1001	$+1.1001 \times 2^1$	+11.001	+3.125
74	0-100-1010	$+1.1010 \times 2^1$	+11.01	+3.25
75	0-100-1011	$+1.1011 \times 2^1$	+11.011	+3.375
76	0-100-1100	$+1.1100 \times 2^1$	+11.1	+3.5
77	0-100-1101	$+1.1101 \times 2^1$	+11.101	+3.625
78	0-100-1110	$+1.1110 \times 2^1$	+11.11	+3.75
79	0-100-1111	$+1.1111 \times 2^1$	+11.111	+3.875
80	0-101-0000	$+1.0000 \times 2^2$	+100.0	+4.0
81	0-101-0001	$+1.0001 \times 2^2$	+100.01	+4.25
82	0-101-0010	$+1.0010 \times 2^2$	+100.1	+4.5
83	0-101-0011	$+1.0011 \times 2^2$	+100.11	+4.75
84	0-101-0100	$+1.0100 \times 2^2$	+101.0	+5.0
85	0-101-0101	$+1.0101 \times 2^2$	+101.01	+5.25
86	0-101-0110	$+1.0110 \times 2^2$	+101.1	+5.5
87	0-101-0111	$+1.0111 \times 2^2$	+101.11	+5.75
88	0-101-1000	$+1.1000 \times 2^2$	+110.0	+6.0
89	0-101-1001	$+1.1001 \times 2^2$	+110.01	+6.25
90	0-101-1010	$+1.1010 \times 2^2$	+110.1	+6.5
91	0-101-1011	$+1.1011 \times 2^2$	+110.11	+6.75
92	0-101-1100	$+1.1100 \times 2^2$	+111.0	+7.0
93	0-101-1101	$+1.1101 \times 2^2$	+111.01	+7.25
94	0-101-1110	$+1.1110 \times 2^2$	+111.1	+7.5
95	0-101-1111	$+1.1111 \times 2^2$	+111.11	+7.75
96	0-110-0000	$+1.0000 \times 2^3$	+1000.0	+8.0
97	0-110-0001	$+1.0001 \times 2^3$	+1000.1	+8.5
98	0-110-0010	$+1.0010 \times 2^3$	+1001.0	+9.0
99	0-110-0011	$+1.0011 \times 2^3$	+1001.1	+9.5
100	0-110-0100	$+1.0100 \times 2^3$	+1010.0	+10.0
101	0-110-0101	$+1.0101 \times 2^3$	+1010.1	+10.5
102	0-110-0110	$+1.0110 \times 2^3$	+1011.0	+11.0
103	0-110-0111	$+1.0111 \times 2^3$	+1011.1	+11.5
104	0-110-1000	$+1.1000 \times 2^3$	+1100.0	+12.0
105	0-110-1001	$+1.1001 \times 2^3$	+1100.1	+12.5
106	0-110-1010	$+1.1010 \times 2^3$	+1101.0	+13.0
107	0-110-1011	$+1.1011 \times 2^3$	+1101.1	+13.5
108	0-110-1100	$+1.1100 \times 2^3$	+1110.0	+14.0
109	0-110-1101	$+1.1101 \times 2^3$	+1110.1	+14.5
110	0-110-1110	$+1.1110 \times 2^3$	+1111.0	+15.0
111	0-110-1111	$+1.1111 \times 2^3$	+1111.1	+15.5
112	0-111-0000	*****	*****	***** +Infinity

113	0-111-0001	*****	*****	*****	+NaN
114	0-111-0010	*****	*****	*****	+NaN
115	0-111-0011	*****	*****	*****	+NaN
116	0-111-0100	*****	*****	*****	+NaN
117	0-111-0101	*****	*****	*****	+NaN
118	0-111-0110	*****	*****	*****	+NaN
119	0-111-0111	*****	*****	*****	+NaN
120	0-111-1000	*****	*****	*****	+NaN
121	0-111-1001	*****	*****	*****	+NaN
122	0-111-1010	*****	*****	*****	+NaN
123	0-111-1011	*****	*****	*****	+NaN
124	0-111-1100	*****	*****	*****	+NaN
125	0-111-1101	*****	*****	*****	+NaN
126	0-111-1110	*****	*****	*****	+NaN
127	0-111-1111	*****	*****	*****	+NaN
128	1-000-0000	-0.0000×2 ⁻²	-0.0	-0.0	Denormalized, -Zero
129	1-000-0001	-0.0001×2 ⁻²	-0.000001	-0.015625	Denormalized
130	1-000-0010	-0.0010×2 ⁻²	-0.00001	-0.03125	Denormalized
131	1-000-0011	-0.0011×2 ⁻²	-0.000011	-0.046875	Denormalized
132	1-000-0100	-0.0100×2 ⁻²	-0.0001	-0.0625	Denormalized
133	1-000-0101	-0.0101×2 ⁻²	-0.000101	-0.078125	Denormalized
134	1-000-0110	-0.0110×2 ⁻²	-0.00011	-0.09375	Denormalized
135	1-000-0111	-0.0111×2 ⁻²	-0.000111	-0.109375	Denormalized
136	1-000-1000	-0.1000×2 ⁻²	-0.001	-0.125	Denormalized
137	1-000-1001	-0.1001×2 ⁻²	-0.001001	-0.140625	Denormalized
138	1-000-1010	-0.1010×2 ⁻²	-0.00101	-0.15625	Denormalized
139	1-000-1011	-0.1011×2 ⁻²	-0.001011	-0.171875	Denormalized
140	1-000-1100	-0.1100×2 ⁻²	-0.0011	-0.1875	Denormalized
141	1-000-1101	-0.1101×2 ⁻²	-0.001101	-0.203125	Denormalized
142	1-000-1110	-0.1110×2 ⁻²	-0.00111	-0.21875	Denormalized
143	1-000-1111	-0.1111×2 ⁻²	-0.001111	-0.234375	Denormalized
144	1-001-0000	-1.0000×2 ⁻²	-0.01	-0.25	
145	1-001-0001	-1.0001×2 ⁻²	-0.010001	-0.265625	
146	1-001-0010	-1.0010×2 ⁻²	-0.01001	-0.28125	
147	1-001-0011	-1.0011×2 ⁻²	-0.010011	-0.296875	
148	1-001-0100	-1.0100×2 ⁻²	-0.0101	-0.3125	
149	1-001-0101	-1.0101×2 ⁻²	-0.010101	-0.328125	
150	1-001-0110	-1.0110×2 ⁻²	-0.01011	-0.34375	
151	1-001-0111	-1.0111×2 ⁻²	-0.010111	-0.359375	
152	1-001-1000	-1.1000×2 ⁻²	-0.011	-0.375	
153	1-001-1001	-1.1001×2 ⁻²	-0.011001	-0.390625	
154	1-001-1010	-1.1010×2 ⁻²	-0.01101	-0.40625	
155	1-001-1011	-1.1011×2 ⁻²	-0.011011	-0.421875	
156	1-001-1100	-1.1100×2 ⁻²	-0.0111	-0.4375	
157	1-001-1101	-1.1101×2 ⁻²	-0.011101	-0.453125	
158	1-001-1110	-1.1110×2 ⁻²	-0.01111	-0.46875	
159	1-001-1111	-1.1111×2 ⁻²	-0.011111	-0.484375	
160	1-010-0000	-1.0000×2 ⁻¹	-0.1	-0.5	
161	1-010-0001	-1.0001×2 ⁻¹	-0.10001	-0.53125	
162	1-010-0010	-1.0010×2 ⁻¹	-0.1001	-0.5625	
163	1-010-0011	-1.0011×2 ⁻¹	-0.10011	-0.59375	
164	1-010-0100	-1.0100×2 ⁻¹	-0.101	-0.625	
165	1-010-0101	-1.0101×2 ⁻¹	-0.10101	-0.65625	
166	1-010-0110	-1.0110×2 ⁻¹	-0.1011	-0.6875	
167	1-010-0111	-1.0111×2 ⁻¹	-0.10111	-0.71875	
168	1-010-1000	-1.1000×2 ⁻¹	-0.11	-0.75	
169	1-010-1001	-1.1001×2 ⁻¹	-0.11001	-0.78125	

170	1-010-1010	-1.1010×2^{-1}	-0.1101	-0.8125
171	1-010-1011	-1.1011×2^{-1}	-0.11011	-0.84375
172	1-010-1100	-1.1100×2^{-1}	-0.111	-0.875
173	1-010-1101	-1.1101×2^{-1}	-0.11101	-0.90625
174	1-010-1110	-1.1110×2^{-1}	-0.1111	-0.9375
175	1-010-1111	-1.1111×2^{-1}	-0.11111	-0.96875
176	1-011-0000	-1.0000×2^0	-1.0	-1.0
177	1-011-0001	-1.0001×2^0	-1.0001	-1.0625
178	1-011-0010	-1.0010×2^0	-1.001	-1.125
179	1-011-0011	-1.0011×2^0	-1.0011	-1.1875
180	1-011-0100	-1.0100×2^0	-1.01	-1.25
181	1-011-0101	-1.0101×2^0	-1.0101	-1.3125
182	1-011-0110	-1.0110×2^0	-1.011	-1.375
183	1-011-0111	-1.0111×2^0	-1.0111	-1.4375
184	1-011-1000	-1.1000×2^0	-1.1	-1.5
185	1-011-1001	-1.1001×2^0	-1.1001	-1.5625
186	1-011-1010	-1.1010×2^0	-1.101	-1.625
187	1-011-1011	-1.1011×2^0	-1.1011	-1.6875
188	1-011-1100	-1.1100×2^0	-1.11	-1.75
189	1-011-1101	-1.1101×2^0	-1.1101	-1.8125
190	1-011-1110	-1.1110×2^0	-1.111	-1.875
191	1-011-1111	-1.1111×2^0	-1.1111	-1.9375
192	1-100-0000	-1.0000×2^1	-10.0	-2.0
193	1-100-0001	-1.0001×2^1	-10.001	-2.125
194	1-100-0010	-1.0010×2^1	-10.01	-2.25
195	1-100-0011	-1.0011×2^1	-10.011	-2.375
196	1-100-0100	-1.0100×2^1	-10.1	-2.5
197	1-100-0101	-1.0101×2^1	-10.101	-2.625
198	1-100-0110	-1.0110×2^1	-10.11	-2.75
199	1-100-0111	-1.0111×2^1	-10.111	-2.875
200	1-100-1000	-1.1000×2^1	-11.0	-3.0
201	1-100-1001	-1.1001×2^1	-11.001	-3.125
202	1-100-1010	-1.1010×2^1	-11.01	-3.25
203	1-100-1011	-1.1011×2^1	-11.011	-3.375
204	1-100-1100	-1.1100×2^1	-11.1	-3.5
205	1-100-1101	-1.1101×2^1	-11.101	-3.625
206	1-100-1110	-1.1110×2^1	-11.11	-3.75
207	1-100-1111	-1.1111×2^1	-11.111	-3.875
208	1-101-0000	-1.0000×2^2	-100.0	-4.0
209	1-101-0001	-1.0001×2^2	-100.01	-4.25
210	1-101-0010	-1.0010×2^2	-100.1	-4.5
211	1-101-0011	-1.0011×2^2	-100.11	-4.75
212	1-101-0100	-1.0100×2^2	-101.0	-5.0
213	1-101-0101	-1.0101×2^2	-101.01	-5.25
214	1-101-0110	-1.0110×2^2	-101.1	-5.5
215	1-101-0111	-1.0111×2^2	-101.11	-5.75
216	1-101-1000	-1.1000×2^2	-110.0	-6.0
217	1-101-1001	-1.1001×2^2	-110.01	-6.25
218	1-101-1010	-1.1010×2^2	-110.1	-6.5
219	1-101-1011	-1.1011×2^2	-110.11	-6.75
220	1-101-1100	-1.1100×2^2	-111.0	-7.0
221	1-101-1101	-1.1101×2^2	-111.01	-7.25
222	1-101-1110	-1.1110×2^2	-111.1	-7.5
223	1-101-1111	-1.1111×2^2	-111.11	-7.75
224	1-110-0000	-1.0000×2^3	-1000.0	-8.0
225	1-110-0001	-1.0001×2^3	-1000.1	-8.5
226	1-110-0010	-1.0010×2^3	-1001.0	-9.0

227	1-110-0011	-1.0011×2 ³	-1001.1	-9.5	
228	1-110-0100	-1.0100×2 ³	-1010.0	-10.0	
229	1-110-0101	-1.0101×2 ³	-1010.1	-10.5	
230	1-110-0110	-1.0110×2 ³	-1011.0	-11.0	
231	1-110-0111	-1.0111×2 ³	-1011.1	-11.5	
232	1-110-1000	-1.1000×2 ³	-1100.0	-12.0	
233	1-110-1001	-1.1001×2 ³	-1100.1	-12.5	
234	1-110-1010	-1.1010×2 ³	-1101.0	-13.0	
235	1-110-1011	-1.1011×2 ³	-1101.1	-13.5	
236	1-110-1100	-1.1100×2 ³	-1110.0	-14.0	
237	1-110-1101	-1.1101×2 ³	-1110.1	-14.5	
238	1-110-1110	-1.1110×2 ³	-1111.0	-15.0	
239	1-110-1111	-1.1111×2 ³	-1111.1	-15.5	
240	1-111-0000	*****	*****	*****	-Infinity
241	1-111-0001	*****	*****	*****	-NaN
242	1-111-0010	*****	*****	*****	-NaN
243	1-111-0011	*****	*****	*****	-NaN
244	1-111-0100	*****	*****	*****	-NaN
245	1-111-0101	*****	*****	*****	-NaN
246	1-111-0110	*****	*****	*****	-NaN
247	1-111-0111	*****	*****	*****	-NaN
248	1-111-1000	*****	*****	*****	-NaN
249	1-111-1001	*****	*****	*****	-NaN
250	1-111-1010	*****	*****	*****	-NaN
251	1-111-1011	*****	*****	*****	-NaN
252	1-111-1100	*****	*****	*****	-NaN
253	1-111-1101	*****	*****	*****	-NaN
254	1-111-1110	*****	*****	*****	-NaN
255	1-111-1111	*****	*****	*****	-NaN