컴퓨터 구조

3장 컴퓨터 산술과 논리 연산3

안형태

anten@kumoh.ac.kr

디지털관 139호

컴퓨터 산술과 논리 연산

6. 부동소수점 수의 표현

부동소수점 수의 표현

- **부동소수점 표현(floating-point representation):** 지수(exponent)를 이용하여 소수점의 위치를 이동시킬 수 있는 수 표현 방법 → 수 표현 범위 확대

- 부동소수점 수(floating-point number)의 일반적인 형태

$$N = (-1)^S M \times B^E$$

  - 단, S: 부호(sign), M: 가수(mantissa), B: 기수(base), E: 지수(exponent)

부동소수점 수의 표현

- 10진 부동소수점 수(decimal floating-point number)
  - **[예]** $274{,}000{,}000{,}000{,}000 \rightarrow 2.74 \times \mathbf{10^{14}}$
    $0.00000000000274 \rightarrow 2.74 \times \mathbf{10^{-12}}$

- 2진 부동소수점 수(binary floating-point number)
  - 기수 $B = 2$
  - **[예]** $11.101 \rightarrow 0.11101 \times \mathbf{2^{2}}$, $0.00001101 \rightarrow 0.1101 \times \mathbf{2^{-4}}$
  - 단일-정밀도(single-precision) 부동소수점 수: **32 비트**
    - C언어에서 float에 활용되는 데이터의 형식
  - 복수-정밀도(double-precision) 부동소수점 수: **64 비트**
    - C언어에서 double에 활용되는 데이터의 형식

단일-정밀도 부동소수점 수 형식

- **[예]** S: 1 비트, E: 8 비트, M : 23 비트

  - 지수(E) 필드의 비트 수 증가 → 표현 가능한 수의 범위(range) 확장
  - 가수(M) 필드의 비트 수 증가 → 정밀도(precision) 증가

- 단일-정밀도 부동소수점의 표현 가능한 수 크기의 범위:
  - $\mathbf{0.5 \times 2^{-128} \sim 0.5 \times 2^{127} \approx 1.47 \times 10^{-39} \sim 1.7 \times 10^{38}}$
  - **[비교]** 32-비트 **고정소수점**(fixed-point) 표현 방식의 경우:
    - $\mathbf{1.0 \times 2^{-31} \sim 1.0 \times 2^{31} \approx 2.0 \times 10^{-9} \sim 2.0 \times 10^{9}}$

같은 수에 대한 부동소수점 표현

- 부동소수점에서 같은 수에 대한 표현이 여러 가지가 존재
  - $\mathbf{0.1101 \times 2^5}$
  - $\mathbf{11.01 \times 2^3}$
  - $\mathbf{0.001101 \times 2^7}$

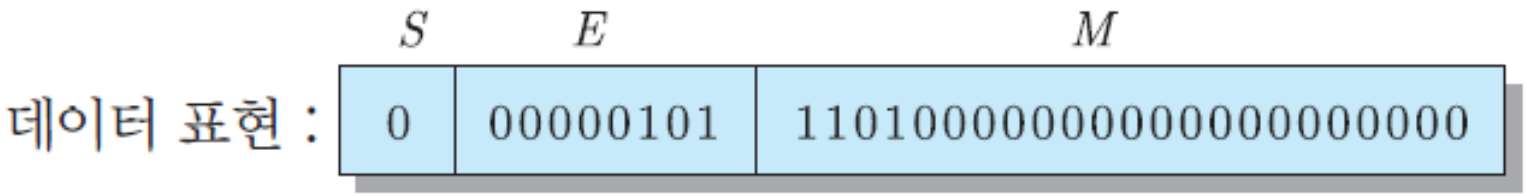
- **정규화된 표현(normalized representation)**
  - 수에 대한 표현을 한 가지로 통일
    - **[예]** 소수점 우측의 첫 번째 비트가 1이 되도록 지수를 조정

$$\pm \mathbf{0.1}bbb..bb \times 2^E$$

  - 위의 예에서 정규화된 표현은 $\mathbf{0.1101 \times 2^5}$

부동소수점 표현

- **[예] 0.1101×2^5**
  - 부호(S) 비트 = 0
  - 지수(E) = 00000101
  - 가수(M) = 1101 0000 0000 0000 0000 000



- 소수점 아래 첫 번째 비트는 항상 1이므로, 반드시 저장할 필요는 없음**(hidden bit)**
  - 가수 23 비트를 이용하여 소수점 아래 24 자리 수까지 표현 가능

바이어스된 지수(biased exponent)

- ‘0’ 표현 문제
  - 부동소수점에서 ‘0’을 표현하려면, 가수는 반드시 0
  - $0 \times 2^E = 0$이므로, 지수는 어떤 값이든 상관 없음
  - 하지만, 부호·가수·지수 비트가 모두 0이 아니면, 0-검사(zero-test) 과정이 복잡 → 이를 단순화하기 위해 **지수를 0으로 만들 방법** 필요
    - 지수의 값이 아주 큰 음수이면, 2^E의 절대값이 거의 0에 가까운 값
- 지수를 **바이어스된 지수**(biased exponent)로 표현
  - 지수에 바이어스 값을 더해 줌
    - [예] 바이어스 127 (excess-127 코드): 지수에 127 (01111111)을 더해 줌
    - [예] 바이어스 128 (excess-128 코드): 지수에 128 (10000000)을 더해 줌
- 용도
  - 0을 표현할 때, 모든 비트가 0이므로 **정수와 동일하게 0-검사 가능**
  - 지수 비트를 부호 없는 8-비트 정수(양의 정수)로 취급할 수 있어서, 부동소수점 수 간 **크기 비교와 정렬이 용이**

바이어스된 지수(biased exponent)

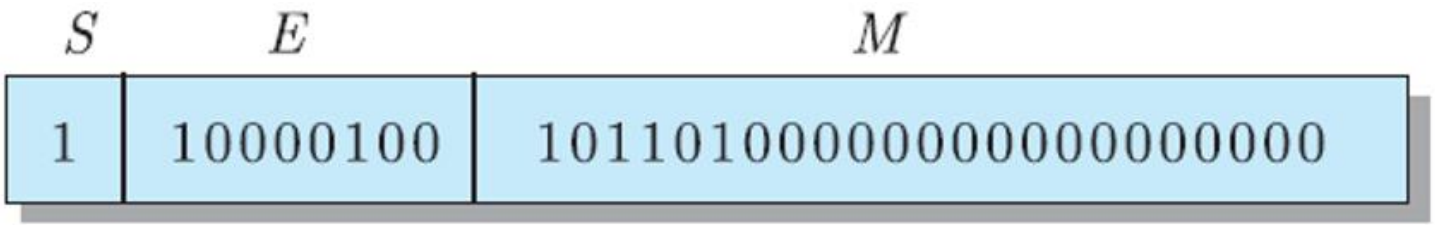
- 8-비트 바이어스된 지수값들(8-bit biased exponents)

지수 비트 패턴	절대값	실제 지수값	
		바이어스 = 127	바이어스 = 128
11111111	255	+128	+127
11111110	254	+127	+126
⋮	⋮	⋮	⋮
10000001	129	+2	+1
10000000	128	+1	0
01111111	127	0	−1
01111110	126	−1	−2
⋮	⋮	⋮	⋮
00000001	1	−126	−127
00000000	0	−127	−128

바이어스된 지수(biased exponent)

- **[예제]** 10 진수 $N = -13.625$에 대한 32비트 부동소수점으로 표현하라. (128 바이어스된 지수 사용)

- **[풀이]**
  - $13.625_{10} = 1101.101_2 = 0.1101101 \times 2^4$
  - 부호(S) 비트 = 1(음수)
  - 지수(E) 비트 = 00000100 + 10000000 = 10000100(바이어스 128을 더함)
  - 가수(M) 비트 = 10110100000000000000000(소수점 우측의 첫번째 '1' 제외)



부동소수점 수의 표현 범위

- 32-비트 부동소수점에서 지수(E)와 가수(M)의 최소값, 최대값

	최소값	최대값
지수(E)	-128	127
가수(M)	0.5	$1 - 2^{-24}$

- 가수(M)의 최소값

  000 0000 0000 0000 0000 0000 → 0.1000 0000 0000 0000 0000 0000

- 가수(M)의 최대값

  1.0000 0000 0000 0000 0000 0000 (=1)

  \- 0.0000 0000 0000 0000 0000 0001 (=$\mathbf{2^{-24}}$)

  0.1111 1111 1111 1111 1111 1111 (=1 - $\mathbf{2^{-24}}$)

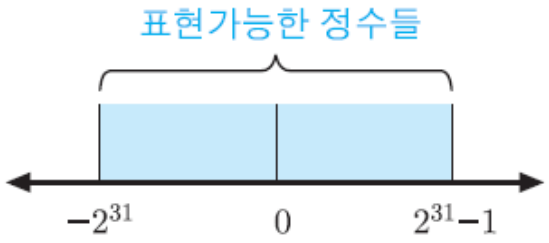
  = 0.999999403953552246093750

부동소수점 수의 표현 범위

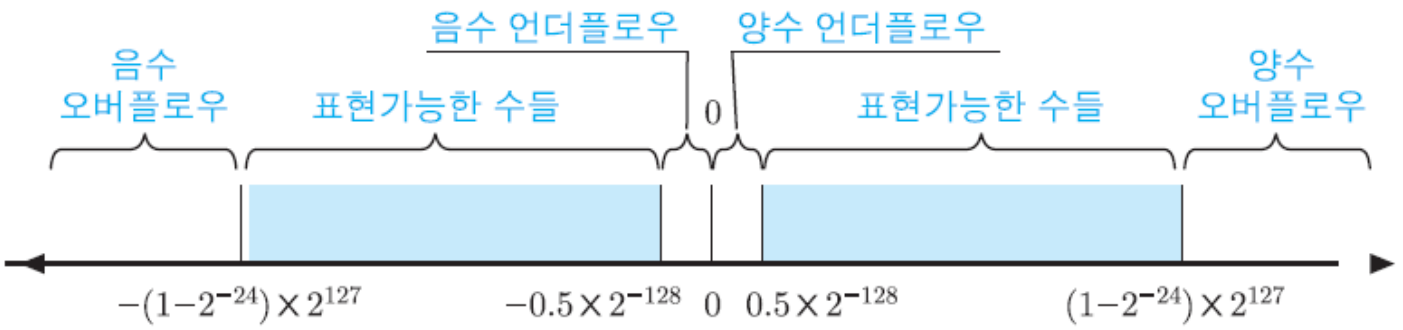
- 부동소수점 수의 표현 범위
  - $\mathbf{0.5 \times 2^{-128}}$ **~** $\mathbf{(1 - 2^{-24}) \times 2^{127}}$ **사이의 양수들**
  - $\mathbf{-(1 - 2^{-24}) \times 2^{127}}$ **~** $\mathbf{-0.5 \times 2^{-128}}$ **사이의 음수들**
  - 수를 표현하는데 필요한 가수가 24-비트를 초과하는 숫자는 제외

- 제외되는 범위
  - $\mathbf{-(1 - 2^{-24}) \times 2^{127}}$ 보다 작은 음수 → **음수 오버플로우(negative overflow)**
  - $\mathbf{-0.5 \times 2^{-128}}$ 보다 큰 음수 → **음수 언더플로우(negative underflow)**
  - $\mathbf{0}$
  - $\mathbf{0.5 \times 2^{-128}}$ 보다 작은 양수 → **양수 언더플로우(positive underflow)**
  - $\mathbf{(1 - 2^{-24}) \times 2^{127}}$ 보다 큰 양수 → **양수 오버플로우(positive overflow)**

32-비트 데이터 형식의 표현 가능한 수의 범위



(a) 2의 보수 정수의 표현 범위



(b) 부동소수점 수의 표현 범위

IEEE 754 표준 부동소수점 수의 형식

- 부동소수점 수의 표현 방식의 통일을 위하여 미국전기전자공학회(IEEE)에서 정의한 국제 표준

- 32-비트 단일-정밀도 부동소수점 수의 표현

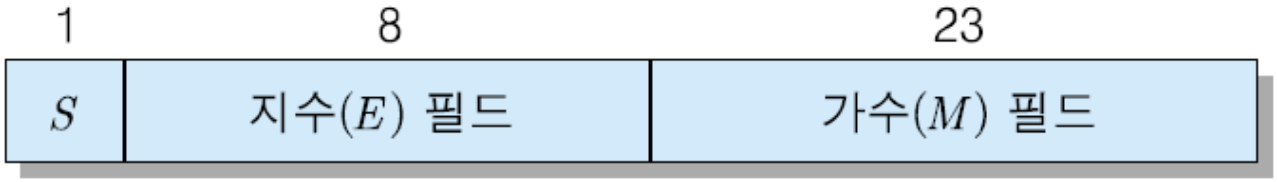
$$N = (-1)^S \times (1.M) \times 2^{E-127}$$

  - 가수: 부호화-크기 표현 사용
  - 지수 필드: **바이어스 127** 사용
  - $1.M \times 2^E$의 형태를 가지며, 소수점 아래의 M 부분만 가수 필드에 저장
    - 소수점 왼쪽의 저장되지 않는 1은 **hidden bit**

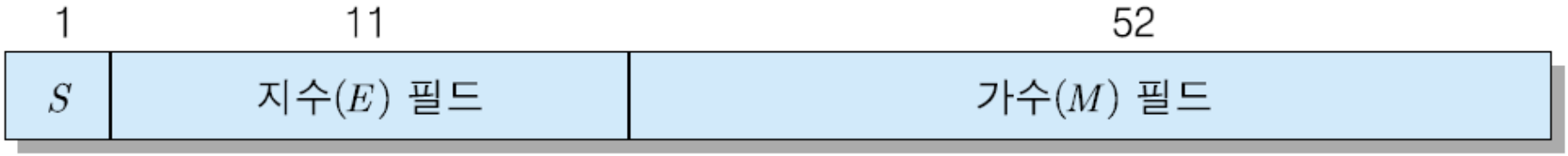
- 64-비트 복수-정밀도 부동소수점 수의 표현

$$N = (-1)^S \times (1.M) \times 2^{E-1023}$$

IEEE 754 표준 부동소수점 수의 형식



(a) 단일-정밀도 형식(single-precision format)

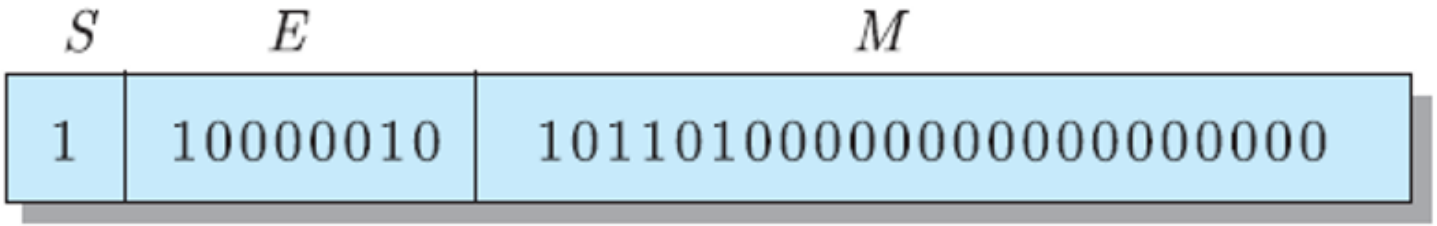


(b) 복수-정밀도 형식(double-precision format)

IEEE 754 표준 부동소수점 수의 형식

- **[예제]** 10 진수 $N = -13.625$를 IEEE 754 단일-정밀도 표준 형식으로 표현하라.

- **[풀이]**
  - $13.625_{10} = 1101.101_2 = 1.101101 \times 2^3$
  - 부호(S) 비트 = 1 (음수)
  - 지수(E) 비트 = 00000011 + 01111111 = 10000010 (바이어스 127을 더함)
  - 가수(M) 비트 = 10110100000000000000000 (소수점 좌측의 첫번째 '1' 제외)

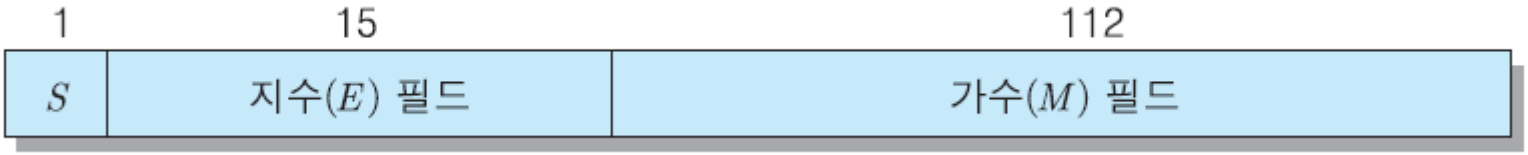


예외(exception) 경우를 포함한 IEEE 754 표준

- 예외 경우를 포함한 정의 (32-비트 형식)
  - 만약 $E = 255$이고 $M \neq 0$이면, N = *Not a Number*(***NaN***)
    - **[TMI]** NaN 발생 조건: 0으로 나누기, 무한대에 0을 곱하기, 무한대에서 무한대로 나누기, 음수에 대한 루트 계산 등
  - 만약 $E = 255$이고 $M = 0$이면, $N = (-1)^S \infty$ → 오버플로우
  - 만약 $0 < E < 255$ 이면, $N = (-1)^S (1.M) 2^{E-127}$
  - 만약 $E = 0$이고 $M \neq 0$ 이면, $N = (-1)^S (0.M) 2^{-126}$ → 언더플로우
  - 만약 $E = 0$이고 $M = 0$이면, $N = (-1)^S 0$ → 0
- 예외 경우를 포함한 정의 (64-비트 형식)
  - 만약 $E = 2047$이고 $M \neq 0$이면, $N = NaN$
  - 만약 $E = 2047$이고 $M = 0$이면, $N = (-1)^S \infty$ → 오버플로우
  - 만약 $0 < E < 2047$ 이면, $N = (-1)^S (1.M) 2^{E-1023}$
  - 만약 $E = 0$이고 $M \neq 0$ 이면, $N = (-1)^S (0.M) 2^{-1022}$ → 언더플로우
  - 만약 $E = 0$이고 $M = 0$이면, $N = (-1)^S 0$ → 0

[TMI] IEEE 754-2008에서 추가된 정의

- 4배수-정밀도(quadruple-precision: 128-비트) 부동소수점 수의 표현



$$N = (-1)^{S}(1.M)2^{E-16383}$$

- 가수: 부호화-크기 표현 사용
- 지수 필드: **바이어스 16383** 사용(범위: -16382 ~ +16383)

[TMI] IEEE 754 형식의 주요 파라미터들

파라미터	부동소수점 형식		
	단일 정밀도	복수 정밀도	4배수 정밀도
비트 수	32	64	128
지수 필드 폭(비트)	8	11	15
가수 필드 폭(비트)	23	52	112
지수 바이어스	127	1023	16383
최대 지수값	127	1023	16383
최소 지수값	−126	−1022	−16382
정수의 대략적 표현 범위 (최소/최대)	$\sim \pm 10^{-38} / \pm 10^{+38}$	$\sim \pm 10^{-308} / \pm 10^{+308}$	$\sim \pm 10^{-4932} / \pm 10^{+4932}$

Q & A

End!