TRUONG DAI HOC KHOA HOC TU NHIEN
KHTN
TP. HO CHI MINH

Floating Point

① Fixed point representation

② Floating point representation

③ IEEE 754 standard

Is it true for all values of x and y:

A. (x>0)||(x-1<0)

B. (x&7)!=7||(x<<29<0)

C. (x*x)>=0

D. x<0||-x<=0

E. x>0||-x>=0

```
Int x=foo();
Int y=bar();
```

Real Numbers

10^m	10^{m-1}		10^2	10^1	10^0		10^{-1}	10^{-2}		10^{-n}
d_m	d_{m-1}	…	d_2	d_1	d_0	.	d_{-1}	d_{-2}	…	d_{-n}
1	9	…	1	0	2	.	0	1	…	6

$$d = \sum_{i=-n}^{m} 10^i \times d_i$$

Binary Representation

2^m	2^{m-1}		2^2	2^1	2^0		2^{-1}	2^{-2}		2^{-n}
b_m	b_{m-1}	…	b_2	b_1	b_0	.	b_{-1}	b_{-2}	…	b_{-n}
1	0	…	1	0	1	.	0	1	…	1

$$b = \sum_{i=-n}^{m} 2^i \times b_i$$

Examples

Fractional value	Binary representation	Decimal representation
1/8	0 . 0 0 1	0.125
25/16	1 . 1 0 0 1	1.5625
43/16	1 0 . 1 0 1 1	2.6875
9/8	1 . 0 0 1	1.125
1/3	0 . (0 1)	0.(3)
1/10		0.1

$$b = \sum_{i=-n}^{m} 2^i \times b_i$$

Precision

8 bits	0.0001100
16 bits	0.0001100110011100
24 bits	0.000110011001100110011100
32 bits	0.00011001100110011001100110011100

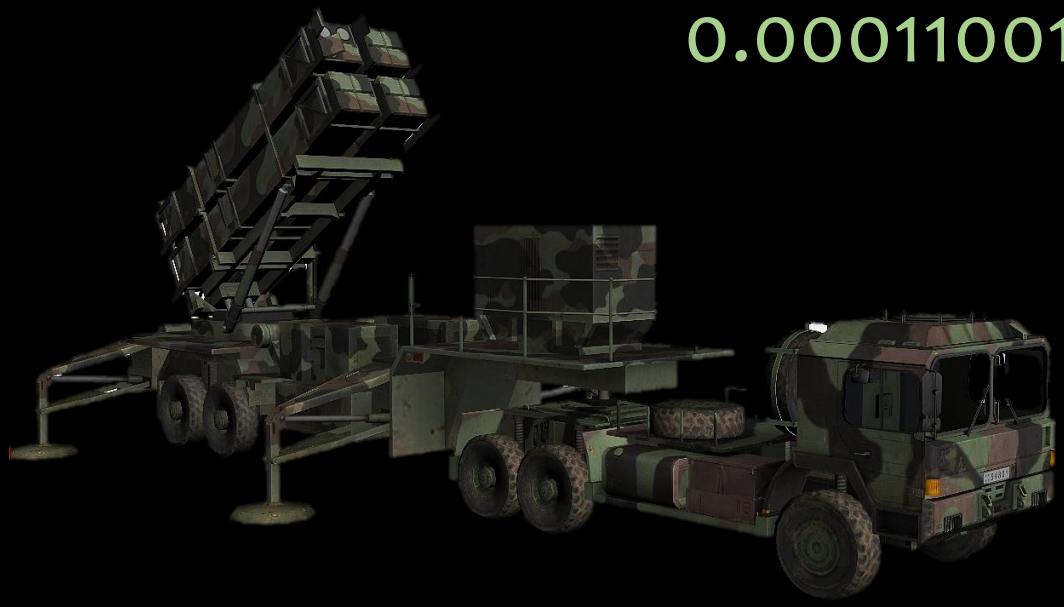
1/10	0.0(0011)	0.1

The Patriot Missile Failure

1,676m/s × 0.34s = 569.84 m

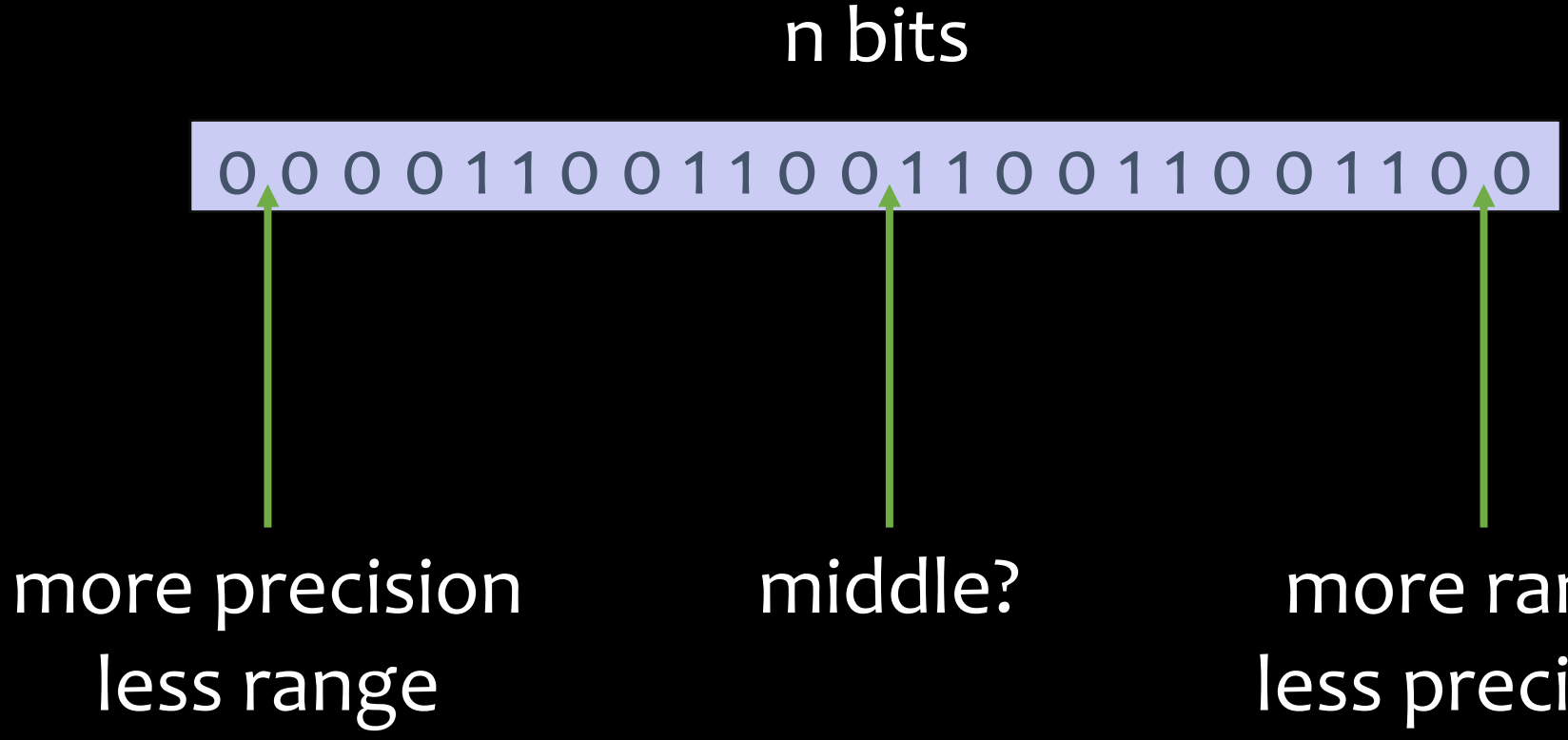
0.000000095s

0.00011001100110011001100



28

Fixed-point representation



Floating-point representation

$$(-1)^S \times M \times 2^E$$

1001.1101 → 1.0011101×2^3

0.011101 → 1.1101×2^{-2}

-0.011101 → $-1 \times 1.1101 \times 2^{-2}$

① Arithmetic formats

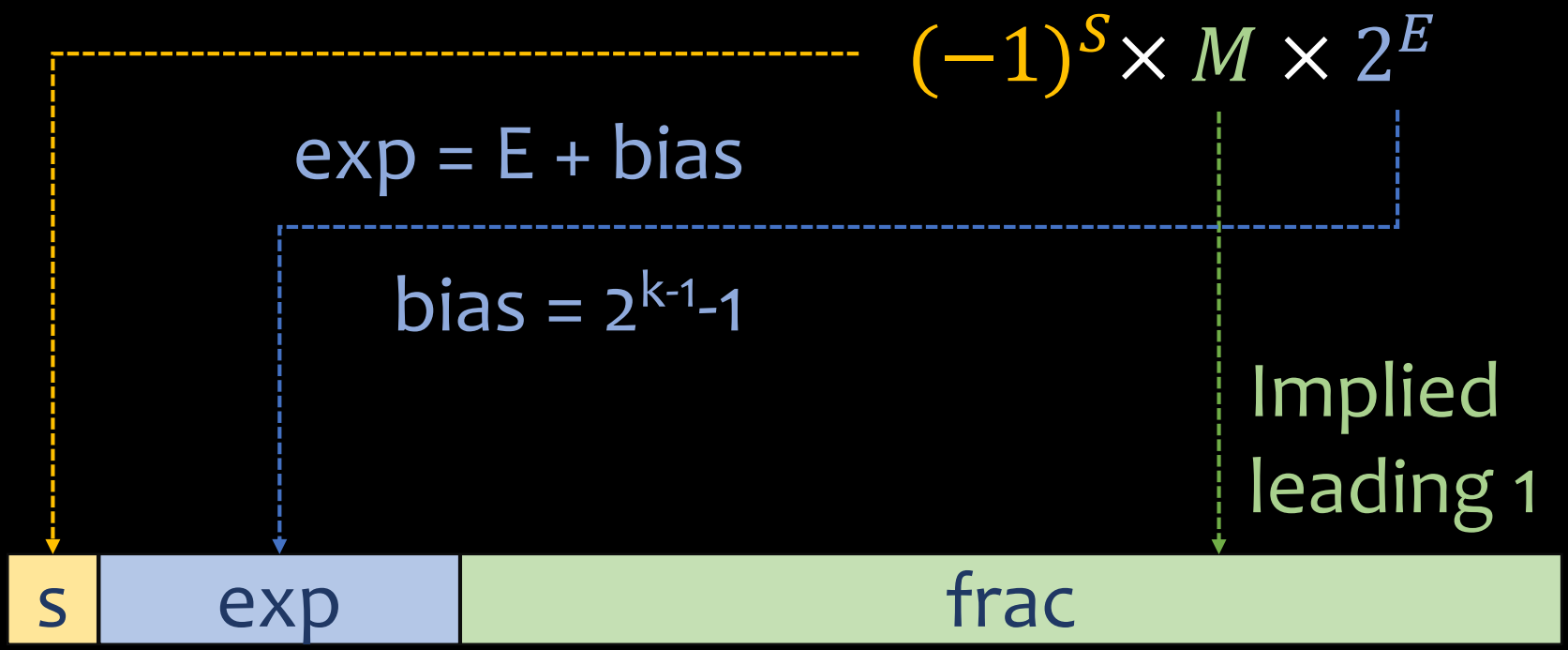
② Interchange formats

③ Rounding rules

④ Operations

⑤ Exception handling

Normalized values



Precision options

Single precision: 32 bits

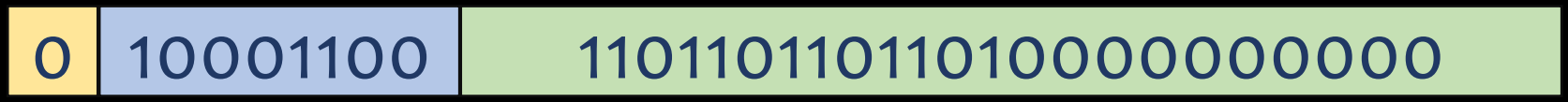
1 8 bits 23 bits

Double precision: 64 bits

1 11 bits 52 bits

Example

Single precision: 32 bits



10001100_2 $11011011011010000000000_2$

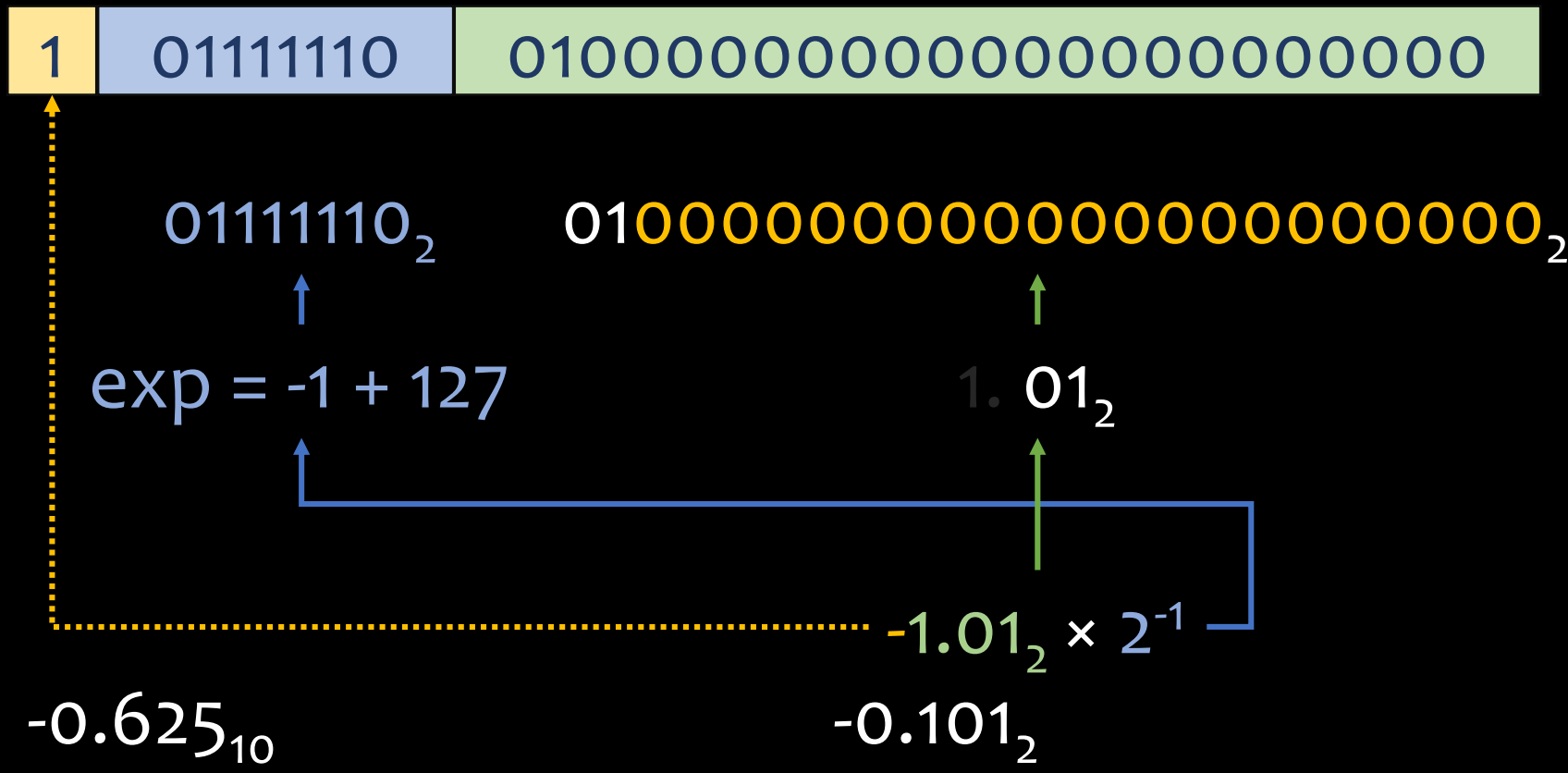
exp = 13 + 127 1.1101101101101_2

$1.1101101101101_2 \times 2^{13}$

15213.0_{10} 1110110110110_2

Example 2

Single precision: 32 bits



Denormalized Values

-0.0 and +0.0

E = 1 - bias

Implied leading 0

Gradual underflow (closest to 0.0)

Special Values

-∞ and +∞

NaN (Not a Number)

Summary

- Fixed-point representation
- Floating-point representation

James Gosling

Java Inventor

“95% of the folks out there are completely clueless about floating-point.”