Computer Arithmetics

Debiprasanna Sahoo
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology Roorkee

Content

Book

Computer Organization and Design: The Hardware/Software Interface-RISC-V Edition, 5th Edition, 2017

Chapter-3

David A. Patterson and John L. Henessey

Reference Books

Computer Organization and Design: The Hardware/Software Interface-MIPS Edition, 5th Edition, 2017

Chapter-3

David A. Patterson and John L. Henessey

Manual

The RISC-V Instruction Set Manual

Volume I: User-Level ISA

Document Version 2.2

Andrew Waterman and Krste Asanovi

*Image from the book and manual unless specified

Primitive Data Types Supported by Programming Languages

Data Types	Example	16 bit System (in Bytes)	32 bit System (in Bytes)	64 bit System (in Bytes)
Character	char c;	1	1	1
Unsigned Character	unsigned char uc;	1	1	1
Integer	int i;	2	4	4
Short Integer	short int s;	2	2	2
Long Integer	long int l;	4	4	8
Unsigned Integer	unsigned int ui;	2	4	4
Unsigned Short	unsigned short us;	2	2	2
Unsigned Long	unsigned long ul;	4	4	8
Long Long	long long ll;	8	8	8
Float	float f;	4	4	4
Double	double d;	8	8	8
Long Double	long double ld;	16	16	16

Characters

Characters are stored as ASCII/Unicode numbers
Range of Signed Characters is -128 to 127

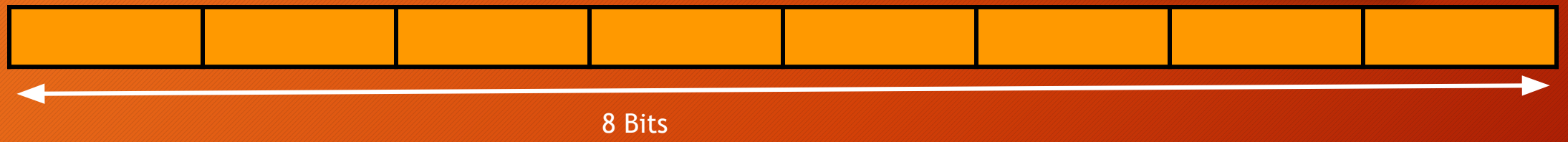
Range of Positive Number Supported

S = 0, 7 bits	0000000 to 1111111
1's Complement	0000000 to 1111111
2's Complement	0000000 to 1111111
Integer	0 to 127

Range of Negative Number Supported

S = 1, 7 bits	0000000 to 1111111
1's Complement	1111111 to 0000000
2's Complement	10000000 to 0000001
Integer	-128 to -1

Unsigned Characters

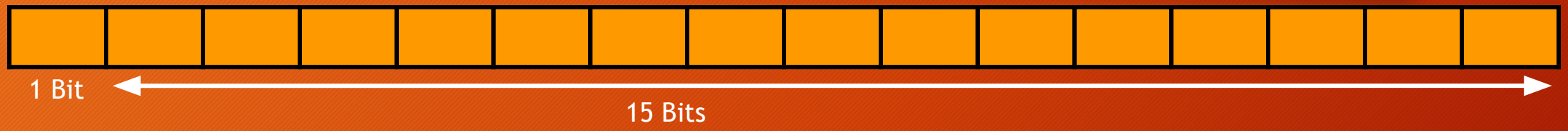


Characters are stored as ASCII numbers
Range of Signed Characters is -128 to 127

Range of Positive Number Supported

8 bits	00000000 to 11111111
1's Complement	00000000 to 11111111
2's Complement	00000000 to 11111111
Integer	0 to 255

Short Integers



Range of Positive Number Supported

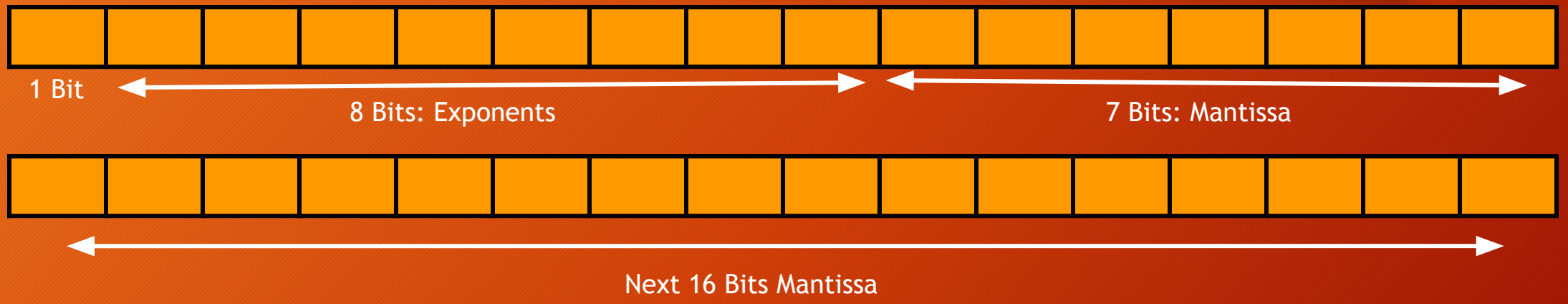
15 bits	000,0000,0000,0000 to	111,1111,1111,1111
1's Complement	000,0000,0000,0000 to	111,1111,1111,1111
2's Complement	000,0000,0000,0000 to	111,1111,1111,1111
Integer	0 to	32767

Range of Negative Number Supported

15 bits	000,0000,0000,0000 to	111,1111,1111,1111
1's Complement	111,1111,1111,1111 to	000,0000,0000,0000
2's Complement	1000,0000,0000,0000 to	000,0000,0000,0001
Integer	-32768 to	-1

Range of Signed Short Integers is -32768 to 32767

Floating Point Numbers



Three parts: Sign bit (S), Exponent (E), Mantissa/Fraction (F).
IEEE 754 Standards represents float as $(-1)^S * F * 2^E$

Normalized Numbers: Number represented in the following binary format of x's and y's where any x and any y can take values 0 or 1

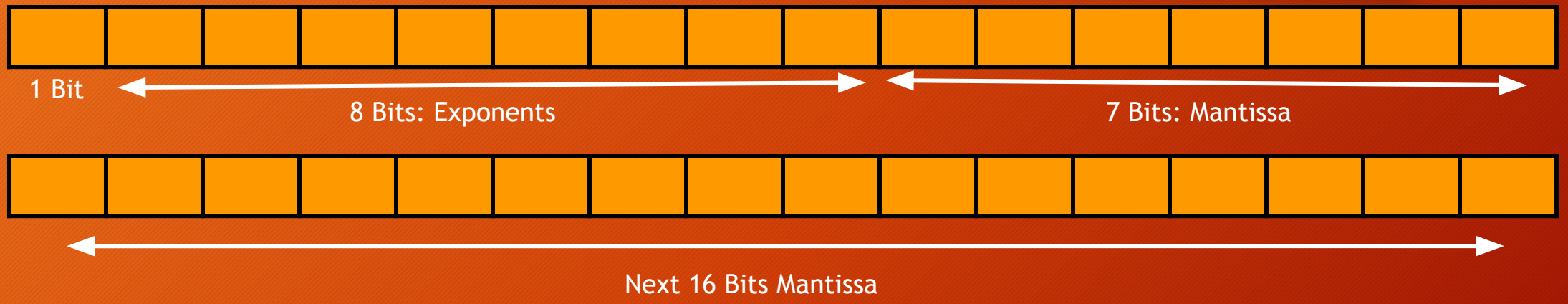
$1.xxxxxxxxxxxxxxxxxxxxxxx * 2^{yyyyyyyy}$

Architectures support both Normalised and De-normalised Numbers

Types of Floating Point Numbers

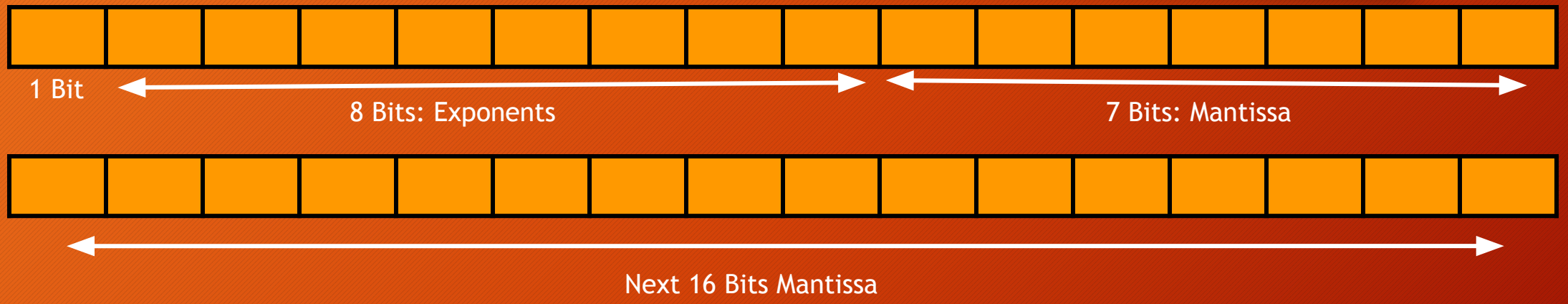
Single Precision			Double Precision			Object
Exponent	Fraction	Range	Exponent	Fraction	Range	
0	0	0	0	0	0	0
0	Non-Zero	2^{-149} to -2^{-149}	0	Non-Zero		+ve or -ve De-normalized Number
1-254	Anything	-1.00000….23 Times * 2^{-126} to 1.11111….23 Times * 2^{127}	1-2046	Anything	-1.00000….52 Times * 2^{-1022} to 1.11111….52 Times * 2^{1023}	+ve or -ve Normalized Number
255	0	+ve or -ve Infinity	2047	0	+ve or -ve Infinity	+ve or -ve Infinity
255	Non-Zero	NaN	255	Non-Zero	NaN	Not a number (NaN)

Normalised Floating Point Numbers



Bias = 127 for Floats and 2046 for Double in the formula fixed by IEEE 754, $(-1)^S * F * 2^{E-Bias}$
Min value of E = 1 and Max value of E = 254
Min value of F = 000000…23 times and Max value of F = 11111…..23 times
When S=0, Max +ve number = 1.11111….23 Times * $2^{254-127}$ = 1.11111….23 Times * 2^{127}
When S=0, Min +ve number = 1.000000….23 Times * 2^{1-127} = 1.00000….23 Times * 2^{-126}
When S=1, Min -ve number = -1.11111….23 Times * $2^{254-127}$ = -1.11111….23 Times * 2^{127}
When S=1, Max -ve number = -1.000000….23 Times * 2^{1-127} = -1.00000….23 Times * 2^{-126}

De-Normalised Floating Point Numbers



Squeeze every bit in the 4/8 Byte to increase the range further.
Bias = 127 for Floats and 2046 for Double in the formula fixed by IEEE 754, $(-1)^{S} * F * 2^{E - Bias}$
Value of E = 0, treat it as 1 for calculations
Min value of F = 000000…22 times…1 and Max value of F = 11111…..23 times
When S=0, Max +ve number = 0.11111….23 Times * 2^{1-127} = 1.11111….22 Times * 2^{-127}
When S=0, Min +ve number = 0.000000….22 Times..1 * 2^{1-127} = 2^{-149}
When S=1, Min -ve number = -0.11111….23 Times * 2^{1-127} = -1.11111….22 Times * 2^{-127}
When S=1, Max -ve number = -0.00000….22 Times..1 * 2^{1-127} = -2^{-149}