

# Floating Point Formats for Machine Learning

IEEE Working Group P3109 interim report  
on 8-bit binary floating-point formats.

24 September 2023

Send questions and comments to the P3109 Secretary <[jeffrey.sarnoff@ieee.org](mailto:jeffrey.sarnoff@ieee.org)>.

Copyright © 2023 by The Institute of Electrical and Electronics Engineers, Incorporated  
Three Park Avenue  
New York, New York 10016-5997, USA  
All rights reserved.

This document is an unapproved draft of a proposed IEEE Standard. As such, this document is subject to change. USE AT YOUR OWN RISK! IEEE copyright statements SHALL NOT BE REMOVED from draft or approved IEEE standards, or modified in any way. Because this is an unapproved draft, this document must not be utilized for any conformance/compliance purposes.

## Table of Contents

1.	Introduction.....	3
2.	Typographical conventions and notation .....	3
3.	Values .....	4
4.	Subnormals.....	5
5.	Not a number (NaN) .....	5
6.	Zero.....	5
7.	Infinities .....	6
8.	Extremal Values .....	6
9.	Classification Operators.....	7
10.	Comparison Operators .....	7
	Appendix A: Numerical Examples .....	9
	Appendix B: Value Tables .....	10
	Value table: binary8p3 .....	11
	Value table: binary8p4 .....	12
	Value table: binary8p5 .....	13

## 1. Introduction

This document represents the results of discussions and decisions made by IEEE working group P3109, “Standard for Arithmetic Formats for Machine Learning”. The Project Authorization Request (PAR) for P3109 defines the scope, need, and stakeholders as follows:

**Scope of proposed standard:** *This standard defines a binary arithmetic and data format for machine learning-optimized domains. It also specifies the default handling of exceptions occurring in this arithmetic. This standard provides a consistent and flexible arithmetic framework optimized for Machine Learning Systems (MLS) in hardware and/or software implementations to minimize the work required to make MLS interoperable with each other as well as other dependent systems. This standard is aligned with the IEEE Std 754-2019 for Floating-Point Arithmetic.*

**Need for the Project:** *Machine Learning Systems have different arithmetic requirements from most other domains. Precisions tend to be lower, and accuracy is measured in dimensions other than just numerical (e.g. inference accuracy). Furthermore, Machine Learning Systems often are integrated into mission-critical and safety-critical systems. With no standards specifically addressing these needs, Machine Learning Systems are built with inconsistent expectations and assumptions that hinder the compatibility and reuse of machine learning hardware, software, and training data.*

**Stakeholders for the Standard:** *System developers, vendors, and users of machine learning applications across many industries and interests including but not limited to compute, storage, medical, telecommunications, e-commerce, fleet-management, automotive, robotics, and security.*

The scope of this interim release is interchange formats only. The working group continues to deliberate on the specification of operations.

## 2. Typographical conventions and notation

**Bold text describes the decisions and specifications of this document.**

Non-bold text is non-normative background material, typically providing rationale and arguments representing the discussions of the WG leading to a decision and specification.

This document specifies 8-bit floating-point interchange formats (binary formats) and associated operations. Binary formats are parameterized by their width, the number of bits spanned in memory (here, 8); and their precision ( $p$ ), the number of bits spanned by the true significand (one more than the number of explicit “mantissa” bits).

**The formats defined herein shall be referred to as “binary8” formats, and further parametrized by precision  $p$  yielding names “binary8pp”.**

For example, “binary8p3” is a format with 3 bits of precision, hence a 2-bit mantissa and a 5-bit exponent field.

### 3. Values

This section describes the set of values that a binary8 format shall represent. The universe of values in existing floating point usage encompasses some finite real numerical values, the nonfinite numerical values positive and negative infinity ( $-\text{Inf}$ ,  $+\text{Inf}$ ), the non-numeric not-a-number values ( $\text{NaN}$ ,  $\text{NaN}_1$ , ...), and negative zero ( $-0$ ). The value set for each binary8 format specifies the set of values that are available in that format.

**Each binary format shall be associated with a unique encoding.** An 8-bit binary encoding is a mapping from 8-bit strings to values. Some of these mappings are included as an Appendix.

**The four special values (0,  $+\text{Inf}$ ,  $-\text{Inf}$ ,  $\text{NaN}$ ) have encodings that are shared by all binary8 formats.**

**Table 1 – Encoding Special Values**

Value	Hexadecimal Encoding	Bit Sequence
Zero	0x00	0000 0000
Positive Infinity ( $+\text{Inf}$ )	0x7F	0111 1111
Negative Infinity ( $-\text{Inf}$ )	0xFF	1111 1111
Not a Number ( $\text{NaN}$ )	0x80	1000 0000

The set of finite floating-point numbers representable with a binary format is determined by two parameters:

- Precision  $p$ , the number of digits in the significand including the implicit leading bit
- Maximum exponent  $emax$ , the exponent of the largest finite value

IEEE-754 2019 includes the radix  $b$  and  $emin$  in the list of format parameters. This document covers binary (radix 2) formats only. The parameter  $emin$  is determinable from other parameters, so is not a format-defining parameter.

**P3109 formats shall define  $emax = 2^{8-p-1} - 1$ .** In IEEE-754,  $emax$  was consistently chosen across formats to be  $2^{w-1} - 1$ , where  $w$  is the exponent field width in bits. This choice has the consequence that the binary8pp value sets are subsets of the IEEE-754 binary16 value set for  $p > 2$ , and a near-symmetric distribution of values below and above the value 1.

**Table 2 – Parameters for Binary Formats \***

Parameter	binary8pp	binary8p5	binary8p4	binary8p3	binary16	binary32	binary64
$k$ , storage width in bits	8	8	8	8	16	32	64
$p$ , precision in bits	$p$	5	4	3	11	24	53
$emax$ , max exponent	$2^{8-p-1} - 1$	3	7	15	15	127	1023
$w$ , exponent field width	$8 - p$	3	4	5	5	8	11
$bias$ , $E - e$	$emax + 1$	4	8	16	15	127	1023
sign bit	1	1	1	1	1	1	1
$t$ , trailing significand field width in bits	$p - 1$	4	3	2	10	23	52

\* Adapted from table 3.5 of IEEE-754 (2019), and extended to include proposed binary8 formats.

## 4. Subnormals

**Binary8 value sets shall include subnormals.**

The IEEE-754 value sets include subnormals. A value with trailing significand field  $m$  and exponent  $e$  is interpreted as  $1. m \times 2^{e-b}$  except when all bits of the exponent bitfield are 0, in which case, the value is  $0. m \times 2^{e-b+1}$ .

When training models, it is common to represent near-zero values for gradients. Subnormal numbers induce equal quantization steps around zero; this expands the reach of binary8 trainable models. In statistical applications, the subnormal range is useful for uniform-like distributions, being uniform around zero. This also supports working with Gaussian-like distributions where numbers around zero are more probable.

## 5. Not a number (NaN)

**Binary8 value sets shall include exactly one NaN, which shall not signal.**

Other floating point formats define several NaN values, denoted (NaN, NaN<sub>1</sub>, ...). NaNs are returned from operations with results outside the set of values. For example, `DIV(0, 0)`, or `ADD(Inf, -Inf)`. Multiple NaN encodings are used in other formats to allow different exceptional conditions to be distinguished.

In the context of machine learning systems, uses of NaN include:

- Debugging of code running on the accelerator. In AI accelerators, exceptions may be difficult or expensive to convey back to user code, so it is common practice to allow NaN values to propagate through calculations in order to indicate that an error has occurred.
- Usage as a “notable value” indicator. In some datasets, for example tabular data, values may be missing. It is useful to use a value outside of the normal numeric range to indicate the position of these values. Particularly where memory usage is a concern, as may be expected in float8 applications, the use of a separate “mask” array, or a list of indices, imposes an extra memory overhead. In some cases, an Inf can be used as a missing value, but given the restricted range of float8, it is likely that infinity shall be used as a separate indicator of “larger/smaller than +/- MaxFloat”.
- The use of multiple NaN payloads is not unknown in statistical code (e.g. the R system has NaN and N/A), but it is not widely used, and in the context of float8, multiple NaNs imposes either additional hardware complexity (using only a subset of the significand range), or a large reduction in encoding space (e.g. 8 codes for E5, 16 codes for E4, 32 codes for E3).

## 6. Zero

**Binary8 formats shall have exactly one zero. This zero value is non-negative.**

The inclusion of negative zero would incur the cost of an additional code point. Given the decision to encode only a single NaN, placing that NaN at the negative zero code point enables the strictly positive and strictly negative number ranges to be symmetric.

A key rationale for inclusion of -0 in IEEE-754 was the consistent implementation of branch cuts in the `atan2` function [Kahan 1987]. Although the `atan` function is common in deep learning, it is used as an activation function, rather than a trigonometric operation, and the `atan2` function is not known in deep learning applications. Furthermore, it is not expected that this standard shall define either `atan` or `atan2`.

A secondary rationale for inclusion of -0 is the hardware simplification offered by its presence in the implementation of sign/magnitude arithmetic. However, the existence of in-market implementations is evidence that the small hardware simplification has not been sufficient to balance the loss of one code point.

It might be considered that the use of integer comparisons in sorting would argue against the placing of NaN at the negative zero code point. For example, the JAX machine learning framework is known to sort using integer comparison [\[link\]](#). However such sorting still requires  $O(n)$  preprocessing and postprocessing steps to enable the use of two's-complement integer comparison, and already has special treatment of NaN and  $-0$ , so eliminating  $-0$  and placing NaN in the  $-0$  position imposes negligible additional burden. As an aside, it is noted that sorting using comparison operations, as typically defined, is undefined in presence of NaN, however the existing practice is to sort NaN (e.g. using `totalOrder`) to the end of the array, and this remains permitted, at no additional cost.

## 7. Infinities

**Binary8 formats shall include positive and negative infinities.**

This decision causes a reduction in dynamic range (252 values rather than 254), but offers improved numerical robustness in important machine learning use cases. Two generic classes of such usage are:

- mask values, for example, in Transformer models in machine learning, [\[ref\]](#).
- representation of overflow.

As illustrated in Appendix A, both usages are facilitated by the presence of infinity.

## 8. Extremal Values

It is practical to offer extremal finite values for supported 8-bit binary interchange formats. Following IEEE 754-2019 naming patterns, we adopt: `maxNormal(T)`, `minNormal(T)`, `minSubnormal(T)` where T is a binary8 format. For example: `maxNormal(binary8p4) == 7/4 * (2^7)`, `minNormal(binary8p5) == 1 / (2^3)`.

	maxNormal	minNormal	minSubnormal
binary8p3	$3/2 * (2^{15})$	$1 / (2^{15})$	$1 / (2^{17})$
binary8p4	$7/4 * (2^7)$	$1 / (2^7)$	$1 / (2^{10})$
binary8p5	$15/8 * (2^3)$	$1 / (2^3)$	$1 / (2^7)$
binary8p6	$31/16 * (2^1)$	$1 / (2^1)$	$1 / (2^6)$

**Table 3: Extremal Values**

## 9. Classification Operators

**Conforming implementations shall provide these classification predicates and the classifier function. The classification predicates and the classifier function shall not signal exceptions.**

<b>predicate</b>	<b>Definition</b>	<b>predicate</b>	<b>definition</b>
isZero	Iff $x$ is 0	isNaN	Iff $x$ is NaN
isInfinite	iff $x$ is infinite	isFinite	Iff $x$ is zero, subnormal or normal
isNormal	Iff $x$ is normal	isSubnormal	Iff $x$ is subnormal
isSignMinus	Iff $x$ has a negative sign <sup>1</sup>		
isCanonical	True <sup>2</sup>	isSignaling	False <sup>3</sup>

**Table 4: Classification Predicates**

<sup>1</sup>isSignMinus(NaN) is True: all binary8 formats encode NaN as 0x80 (0b10000000).

<sup>2</sup>There are no non-canonical binary8 interchange formats.

<sup>3</sup>All binary8 formats have one NaN; it does not signal.

### *The Classifier function*

```
enum class(x)
    NaN
    Zero
    positiveInfinity
    positiveNormal
    positiveSubnormal
    negativeInfinity
    negativeNormal
    negativeSubnormal
end
```

## 10. Comparison Operators

**Conforming implementations shall provide these comparison operators and the totalOrder(  $x$ ,  $y$  ) function.**

Comparison operators are two argument predicates and their negations that return { True, False }. Comparisons shall not raise exceptions. Comparisons are either ordered or unordered. A comparison is unordered iff either argument is NaN. All other comparisons are ordered.

For { =, >, ≥, <, ≤, ≡ }, if any argument is NaN the result is False.

For { ≠, >, ≥, <, ≤, ≡ }, if any argument is NaN the result is True.

Otherwise, the result of a comparison shall match the mathematical result.

**Table 5: Comparison Predicates and Negations**

<b>math symbol</b>	<b>predicate <i>true relations</i></b>	<b>math symbol</b>	<b>negation <i>true relations</i></b>
=	CompareEqual <i>equal</i>	≠, NOT =	CompareNotEqual <i>less than, greater than, unordered</i>
>	CompareGreater <i>greater than</i>	≠, NOT >	CompareNotGreater <i>less than, equal, unordered</i>
≥	CompareGreaterEqual <i>equal, greater than</i>	≠, NOT ≥	CompareLessUnordered <i>less than, unordered</i>
<	CompareLess <i>less than</i>	≠, NOT <	CompareNotLess <i>greater than, equal, unordered</i>
≤	CompareLessEqual <i>less than, equal</i>	≠, NOT ≤	CompareGreaterUnordered <i>greater than, unordered</i>
≡	CompareOrdered <i>less than, equal, greater than</i>	≠, NOT ≡	CompareUnordered <i>unordered</i>

#### *The totalOrder predicate*

totalOrder( *x*, *y* ) provides a total ordering over each binary8 format's value set. It shall not raise any exceptions. totalOrder( *x*, *y* ) shall return { True, False } in accord with the logic given below.

```

boolean totalOrder( x, y )
  if ! ( isNaN(x) || isNaN(y) )
    return compareLessEqual( x, y )
  else
    return isNaN( x )1
  end
end

```

<sup>1</sup>All binary8 formats encode NaN as 0x80. The most significant bit is set, so, following 754, it is as -NaN.

Logical operations used within totalOrder()

- ! is the logical negation operator: !true == false, !false == true.
- || is the short-circuiting, left-associative logical OR.
- if *a* is true, *a* || *b* returns true without evaluating *b*.
- if *a* evaluates as false, *a* || *b* returns the evaluation of *b*.
- ( *a* || *b* || *c* ) evaluates as ( *a* || *b* ) || *c*.



## Appendix A: Numerical Examples

### Mask Values

A common use for  $\infty$  is to create masks, for example, in Transformer models in machine learning, [ref]. These values, assembled in mask matrix  $M$  with values  $M_{ij} \in \{0, -\infty\}$  are typically be added to computed values  $A$ , in a computation such as:

$$\log\left(\sum\left(\exp(\tau * (A + M))\right)\right)$$

where  $\tau$  is a “temperature” or “base” parameter [ref]. This calculation depends on the property that  $\exp(\tau * A_{ij} - \infty) = 0$ . It is clear that where  $M_{ij}$  is a large float (e.g. 480), then  $\exp(-480)$  is an extremely small number, clearly much closer to zero than to any other value. However, careful implementations do not execute the calculation as written, and instead fuse the  $\log(\sum(\exp(v)))$  operation into a single operation  $\text{logsumexp}(v)$ , whose implementation makes use of the identity

$$\text{logsumexp}(v) = \text{logsumexp}(v - \max(v)) + \max(v)$$

Without the “sticky” properties of Inf, this would produce incorrect answers. For example, in a format where MaxFloat=224 without Inf, and MaxFloat=240 with Inf:

$$\text{logsumexp}(\tau * [-224, -\infty]) \rightarrow \text{logsumexp}(\tau * [0, -\infty])$$

while

$$\text{logsumexp}(\tau * [-224, -240]) \rightarrow \text{logsumexp}(\tau * [0, -16])$$

If  $\tau = 1$  and all calculations are done in 8-bit floating point, then the answer will be the same, as  $\exp(-16) = 0$ , but if  $\tau$  is small, or calculations are done in mixed precision, as is common with 8-bit floating point, the loss of “stickiness” shall silently yield unexpected answers. It is not expected that the full calculation shall be done in 8-bit floating point, but the subtraction of the maximum value (and computation of the maximum) might reasonably be in 8-bit floating point.

### Overflow to Infinity

A second use of infinity is to indicate overflow on conversion to the binary8 type. Existing implementations offer several behaviours on overflow: overflow to infinity, saturation to MaxFloat, and overflow to NaN. The existence of a code point for infinity allows any of these options to be implemented in a given instantiation, while removing the code point removes the possibility of implementing the first.

## Appendix B: Value Tables

Value tables mapping 8-bit strings to value sets are provided in this section.

A typical entry is of the form:

HEX    BINARY            = BINARY\_FLOAT    = DECIMAL  
0x01 0\_00000\_01 = +0b0.01\*2<sup>-15</sup> = 7.62939453125e-06

Where the fields are interpreted as follows:

HEX	Hexadecimal encoding of the code point
BINARY	Binary expansion of the code point, with underscores separating sign_exponent_significand
BINARY_FLOAT	The precise float value as a binary fraction followed by 2 <sup>e</sup> with decimal exponent <i>e</i>
DECIMAL	The decimal expansion of the value

## Value table: binary8p3

0x00 = 0.00000_00 = +0b0.00*2 <sup>-15</sup> = 0	0x40 = 0.10000_00 = +0b1.00*2 <sup>0</sup> = 1	0x80 = 1.00000_00 = NaN	0xc0 = 1.10000_00 = -0b1.00*2 <sup>0</sup> = -1
0x01 = 0.00000_01 = +0b0.01*2 <sup>-15</sup> = 7.62939e-06	0x41 = 0.10000_01 = +0b1.01*2 <sup>0</sup> = 1.25	0x81 = 1.00000_01 = -0b0.01*2 <sup>-15</sup> = -7.62939e-06	0xc1 = 1.10000_01 = -0b1.01*2 <sup>0</sup> = -1.25
0x02 = 0.00000_10 = +0b0.10*2 <sup>-15</sup> = 1.52588e-05	0x42 = 0.10000_10 = +0b1.10*2 <sup>0</sup> = 1.5	0x82 = 1.00000_10 = -0b0.10*2 <sup>-15</sup> = -1.52588e-05	0xc2 = 1.10000_10 = -0b1.10*2 <sup>0</sup> = -1.5
0x03 = 0.00000_11 = +0b0.11*2 <sup>-15</sup> = 2.28882e-05	0x43 = 0.10000_11 = +0b1.11*2 <sup>0</sup> = 1.75	0x83 = 1.00000_11 = -0b0.11*2 <sup>-15</sup> = -2.28882e-05	0xc3 = 1.10000_11 = -0b1.11*2 <sup>0</sup> = -1.75
0x04 = 0.00001_00 = +0b1.00*2 <sup>-14</sup> = 3.05176e-05	0x44 = 0.10001_00 = +0b1.00*2 <sup>1</sup> = 2	0x84 = 1.00001_00 = -0b1.00*2 <sup>-14</sup> = -3.05176e-05	0xc4 = 1.10001_00 = -0b1.00*2 <sup>1</sup> = -2
0x05 = 0.00001_01 = +0b1.01*2 <sup>-14</sup> = 3.8147e-05	0x45 = 0.10001_01 = +0b1.01*2 <sup>1</sup> = 2.5	0x85 = 1.00001_01 = -0b1.01*2 <sup>-14</sup> = -3.8147e-05	0xc5 = 1.10001_01 = -0b1.01*2 <sup>1</sup> = -2.5
0x06 = 0.00001_10 = +0b1.10*2 <sup>-14</sup> = 4.57764e-05	0x46 = 0.10001_10 = +0b1.10*2 <sup>1</sup> = 3	0x86 = 1.00001_10 = -0b1.10*2 <sup>-14</sup> = -4.57764e-05	0xc6 = 1.10001_10 = -0b1.10*2 <sup>1</sup> = -3
0x07 = 0.00001_11 = +0b1.11*2 <sup>-14</sup> = 5.34058e-05	0x47 = 0.10001_11 = +0b1.11*2 <sup>1</sup> = 3.5	0x87 = 1.00001_11 = -0b1.11*2 <sup>-14</sup> = -5.34058e-05	0xc7 = 1.10001_11 = -0b1.11*2 <sup>1</sup> = -3.5
0x08 = 0.00010_00 = +0b1.00*2 <sup>-14</sup> = 6.10352e-05	0x48 = 0.10010_00 = +0b1.00*2 <sup>2</sup> = 4	0x88 = 1.00010_00 = -0b1.00*2 <sup>-14</sup> = -6.10352e-05	0xc8 = 1.10010_00 = -0b1.00*2 <sup>2</sup> = -4
0x09 = 0.00010_01 = +0b1.01*2 <sup>-14</sup> = 7.62939e-05	0x49 = 0.10010_01 = +0b1.01*2 <sup>2</sup> = 5	0x89 = 1.00010_01 = -0b1.01*2 <sup>-14</sup> = -7.62939e-05	0xc9 = 1.10010_01 = -0b1.01*2 <sup>2</sup> = -5
0x0a = 0.00010_10 = +0b1.10*2 <sup>-14</sup> = 9.15527e-05	0x4a = 0.10010_10 = +0b1.10*2 <sup>2</sup> = 6	0x8a = 1.00010_10 = -0b1.10*2 <sup>-14</sup> = -9.15527e-05	0xca = 1.10010_10 = -0b1.10*2 <sup>2</sup> = -6
0x0b = 0.00010_11 = +0b1.11*2 <sup>-14</sup> = 0.000106812	0x4b = 0.10010_11 = +0b1.11*2 <sup>2</sup> = 7	0x8b = 1.00010_11 = -0b1.11*2 <sup>-14</sup> = -0.000106812	0xcb = 1.10010_11 = -0b1.11*2 <sup>2</sup> = -7
0x0c = 0.00011_00 = +0b1.00*2 <sup>-13</sup> = 0.00012207	0x4c = 0.10011_00 = +0b1.00*2 <sup>3</sup> = 8	0x8c = 1.00011_00 = -0b1.00*2 <sup>-13</sup> = -0.00012207	0xcc = 1.10011_00 = -0b1.00*2 <sup>3</sup> = -8
0x0d = 0.00011_01 = +0b1.01*2 <sup>-13</sup> = 0.000152588	0x4d = 0.10011_01 = +0b1.01*2 <sup>3</sup> = 10	0x8d = 1.00011_01 = -0b1.01*2 <sup>-13</sup> = -0.000152588	0xcd = 1.10011_01 = -0b1.01*2 <sup>3</sup> = -10
0x0e = 0.00011_10 = +0b1.10*2 <sup>-13</sup> = 0.000183105	0x4e = 0.10011_10 = +0b1.10*2 <sup>3</sup> = 12	0x8e = 1.00011_10 = -0b1.10*2 <sup>-13</sup> = -0.000183105	0xce = 1.10011_10 = -0b1.10*2 <sup>3</sup> = -12
0x0f = 0.00011_11 = +0b1.11*2 <sup>-13</sup> = 0.000213623	0x4f = 0.10011_11 = +0b1.11*2 <sup>3</sup> = 14	0x8f = 1.00011_11 = -0b1.11*2 <sup>-13</sup> = -0.000213623	0xcf = 1.10011_11 = -0b1.11*2 <sup>3</sup> = -14
0x10 = 0.00100_00 = +0b1.00*2 <sup>-12</sup> = 0.000244141	0x50 = 0.10100_00 = +0b1.00*2 <sup>4</sup> = 16	0x90 = 1.00100_00 = -0b1.00*2 <sup>-12</sup> = -0.000244141	0xd0 = 1.10100_00 = -0b1.00*2 <sup>4</sup> = -16
0x11 = 0.00100_01 = +0b1.01*2 <sup>-12</sup> = 0.000305176	0x51 = 0.10100_01 = +0b1.01*2 <sup>4</sup> = 20	0x91 = 1.00100_01 = -0b1.01*2 <sup>-12</sup> = -0.000305176	0xd1 = 1.10100_01 = -0b1.01*2 <sup>4</sup> = -20
0x12 = 0.00100_10 = +0b1.10*2 <sup>-12</sup> = 0.000366211	0x52 = 0.10100_10 = +0b1.10*2 <sup>4</sup> = 24	0x92 = 1.00100_10 = -0b1.10*2 <sup>-12</sup> = -0.000366211	0xd2 = 1.10100_10 = -0b1.10*2 <sup>4</sup> = -24
0x13 = 0.00100_11 = +0b1.11*2 <sup>-12</sup> = 0.000427246	0x53 = 0.10100_11 = +0b1.11*2 <sup>4</sup> = 28	0x93 = 1.00100_11 = -0b1.11*2 <sup>-12</sup> = -0.000427246	0xd3 = 1.10100_11 = -0b1.11*2 <sup>4</sup> = -28
0x14 = 0.00101_00 = +0b1.00*2 <sup>-11</sup> = 0.000488281	0x54 = 0.10101_00 = +0b1.00*2 <sup>5</sup> = 32	0x94 = 1.00101_00 = -0b1.00*2 <sup>-11</sup> = -0.000488281	0xd4 = 1.10101_00 = -0b1.00*2 <sup>5</sup> = -32
0x15 = 0.00101_01 = +0b1.01*2 <sup>-11</sup> = 0.000610352	0x55 = 0.10101_01 = +0b1.01*2 <sup>5</sup> = 40	0x95 = 1.00101_01 = -0b1.01*2 <sup>-11</sup> = -0.000610352	0xd5 = 1.10101_01 = -0b1.01*2 <sup>5</sup> = -40
0x16 = 0.00101_10 = +0b1.10*2 <sup>-11</sup> = 0.000732422	0x56 = 0.10101_10 = +0b1.10*2 <sup>5</sup> = 48	0x96 = 1.00101_10 = -0b1.10*2 <sup>-11</sup> = -0.000732422	0xd6 = 1.10101_10 = -0b1.10*2 <sup>5</sup> = -48
0x17 = 0.00101_11 = +0b1.11*2 <sup>-11</sup> = 0.000854492	0x57 = 0.10101_11 = +0b1.11*2 <sup>5</sup> = 56	0x97 = 1.00101_11 = -0b1.11*2 <sup>-11</sup> = -0.000854492	0xd7 = 1.10101_11 = -0b1.11*2 <sup>5</sup> = -56
0x18 = 0.00110_00 = +0b1.00*2 <sup>-10</sup> = 0.000976562	0x58 = 0.10110_00 = +0b1.00*2 <sup>6</sup> = 64	0x98 = 1.00110_00 = -0b1.00*2 <sup>-10</sup> = -0.000976562	0xd8 = 1.10110_00 = -0b1.00*2 <sup>6</sup> = -64
0x19 = 0.00110_01 = +0b1.01*2 <sup>-10</sup> = 0.0012207	0x59 = 0.10110_01 = +0b1.01*2 <sup>6</sup> = 80	0x99 = 1.00110_01 = -0b1.01*2 <sup>-10</sup> = -0.0012207	0xd9 = 1.10110_01 = -0b1.01*2 <sup>6</sup> = -80
0x1a = 0.00110_10 = +0b1.10*2 <sup>-10</sup> = 0.00146484	0x5a = 0.10110_10 = +0b1.10*2 <sup>6</sup> = 96	0x9a = 1.00110_10 = -0b1.10*2 <sup>-10</sup> = -0.00146484	0xda = 1.10110_10 = -0b1.10*2 <sup>6</sup> = -96
0x1b = 0.00110_11 = +0b1.11*2 <sup>-10</sup> = 0.00170898	0x5b = 0.10110_11 = +0b1.11*2 <sup>6</sup> = 112	0x9b = 1.00110_11 = -0b1.11*2 <sup>-10</sup> = -0.00170898	0xdb = 1.10110_11 = -0b1.11*2 <sup>6</sup> = -112
0x1c = 0.00111_00 = +0b1.00*2 <sup>-9</sup> = 0.00195312	0x5c = 0.10111_00 = +0b1.00*2 <sup>7</sup> = 128	0x9c = 1.00111_00 = -0b1.00*2 <sup>-9</sup> = -0.00195312	0xdc = 1.10111_00 = -0b1.00*2 <sup>7</sup> = -128
0x1d = 0.00111_01 = +0b1.01*2 <sup>-9</sup> = 0.00244141	0x5d = 0.10111_01 = +0b1.01*2 <sup>7</sup> = 160	0x9d = 1.00111_01 = -0b1.01*2 <sup>-9</sup> = -0.00244141	0xdd = 1.10111_01 = -0b1.01*2 <sup>7</sup> = -160
0x1e = 0.00111_10 = +0b1.10*2 <sup>-9</sup> = 0.00292969	0x5e = 0.10111_10 = +0b1.10*2 <sup>7</sup> = 192	0x9e = 1.00111_10 = -0b1.10*2 <sup>-9</sup> = -0.00292969	0xde = 1.10111_10 = -0b1.10*2 <sup>7</sup> = -192
0x1f = 0.00111_11 = +0b1.11*2 <sup>-9</sup> = 0.00341797	0x5f = 0.10111_11 = +0b1.11*2 <sup>7</sup> = 224	0x9f = 1.00111_11 = -0b1.11*2 <sup>-9</sup> = -0.00341797	0xdf = 1.10111_11 = -0b1.11*2 <sup>7</sup> = -224
0x20 = 0.01000_00 = +0b1.00*2 <sup>-8</sup> = 0.00390625	0x60 = 0.11000_00 = +0b1.00*2 <sup>8</sup> = 256	0xa0 = 1.01000_00 = -0b1.00*2 <sup>-8</sup> = -0.00390625	0xe0 = 1.11000_00 = -0b1.00*2 <sup>8</sup> = -256
0x21 = 0.01000_01 = +0b1.01*2 <sup>-8</sup> = 0.00488281	0x61 = 0.11000_01 = +0b1.01*2 <sup>8</sup> = 320	0xa1 = 1.01000_01 = -0b1.01*2 <sup>-8</sup> = -0.00488281	0xe1 = 1.11000_01 = -0b1.01*2 <sup>8</sup> = -320
0x22 = 0.01000_10 = +0b1.10*2 <sup>-8</sup> = 0.00585938	0x62 = 0.11000_10 = +0b1.10*2 <sup>8</sup> = 384	0xa2 = 1.01000_10 = -0b1.10*2 <sup>-8</sup> = -0.00585938	0xe2 = 1.11000_10 = -0b1.10*2 <sup>8</sup> = -384
0x23 = 0.01000_11 = +0b1.11*2 <sup>-8</sup> = 0.00683594	0x63 = 0.11000_11 = +0b1.11*2 <sup>8</sup> = 448	0xa3 = 1.01000_11 = -0b1.11*2 <sup>-8</sup> = -0.00683594	0xe3 = 1.11000_11 = -0b1.11*2 <sup>8</sup> = -448
0x24 = 0.01001_00 = +0b1.00*2 <sup>-7</sup> = 0.0078125	0x64 = 0.11001_00 = +0b1.00*2 <sup>9</sup> = 512	0xa4 = 1.01001_00 = -0b1.00*2 <sup>-7</sup> = -0.0078125	0xe4 = 1.11001_00 = -0b1.00*2 <sup>9</sup> = -512
0x25 = 0.01001_01 = +0b1.01*2 <sup>-7</sup> = 0.00976562	0x65 = 0.11001_01 = +0b1.01*2 <sup>9</sup> = 640	0xa5 = 1.01001_01 = -0b1.01*2 <sup>-7</sup> = -0.00976562	0xe5 = 1.11001_01 = -0b1.01*2 <sup>9</sup> = -640
0x26 = 0.01001_10 = +0b1.10*2 <sup>-7</sup> = 0.0117188	0x66 = 0.11001_10 = +0b1.10*2 <sup>9</sup> = 768	0xa6 = 1.01001_10 = -0b1.10*2 <sup>-7</sup> = -0.0117188	0xe6 = 1.11001_10 = -0b1.10*2 <sup>9</sup> = -768
0x27 = 0.01001_11 = +0b1.11*2 <sup>-7</sup> = 0.0136719	0x67 = 0.11001_11 = +0b1.11*2 <sup>9</sup> = 896	0xa7 = 1.01001_11 = -0b1.11*2 <sup>-7</sup> = -0.0136719	0xe7 = 1.11001_11 = -0b1.11*2 <sup>9</sup> = -896
0x28 = 0.01010_00 = +0b1.00*2 <sup>-6</sup> = 0.015625	0x68 = 0.11010_00 = +0b1.00*2 <sup>10</sup> = 1024	0xa8 = 1.01010_00 = -0b1.00*2 <sup>-6</sup> = -0.015625	0xe8 = 1.11010_00 = -0b1.00*2 <sup>10</sup> = -1024
0x29 = 0.01010_01 = +0b1.01*2 <sup>-6</sup> = 0.0195312	0x69 = 0.11010_01 = +0b1.01*2 <sup>10</sup> = 1280	0xa9 = 1.01010_01 = -0b1.01*2 <sup>-6</sup> = -0.0195312	0xe9 = 1.11010_01 = -0b1.01*2 <sup>10</sup> = -1280
0x2a = 0.01010_10 = +0b1.10*2 <sup>-6</sup> = 0.0234375	0x6a = 0.11010_10 = +0b1.10*2 <sup>10</sup> = 1536	0xaa = 1.01010_10 = -0b1.10*2 <sup>-6</sup> = -0.0234375	0xea = 1.11010_10 = -0b1.10*2 <sup>10</sup> = -1536
0x2b = 0.01010_11 = +0b1.11*2 <sup>-6</sup> = 0.0273438	0x6b = 0.11010_11 = +0b1.11*2 <sup>10</sup> = 1792	0xab = 1.01010_11 = -0b1.11*2 <sup>-6</sup> = -0.0273438	0xeb = 1.11010_11 = -0b1.11*2 <sup>10</sup> = -1792
0x2c = 0.01011_00 = +0b1.00*2 <sup>-5</sup> = 0.03125	0x6c = 0.11011_00 = +0b1.00*2 <sup>11</sup> = 2048	0xac = 1.01011_00 = -0b1.00*2 <sup>-5</sup> = -0.03125	0xec = 1.11011_00 = -0b1.00*2 <sup>11</sup> = -2048
0x2d = 0.01011_01 = +0b1.01*2 <sup>-5</sup> = 0.0390625	0x6d = 0.11011_01 = +0b1.01*2 <sup>11</sup> = 2560	0xad = 1.01011_01 = -0b1.01*2 <sup>-5</sup> = -0.0390625	0xed = 1.11011_01 = -0b1.01*2 <sup>11</sup> = -2560
0x2e = 0.01011_10 = +0b1.10*2 <sup>-5</sup> = 0.046875	0x6e = 0.11011_10 = +0b1.10*2 <sup>11</sup> = 3072	0xae = 1.01011_10 = -0b1.10*2 <sup>-5</sup> = -0.046875	0xee = 1.11011_10 = -0b1.10*2 <sup>11</sup> = -3072
0x2f = 0.01011_11 = +0b1.11*2 <sup>-5</sup> = 0.0546875	0x6f = 0.11011_11 = +0b1.11*2 <sup>11</sup> = 3584	0xaf = 1.01011_11 = -0b1.11*2 <sup>-5</sup> = -0.0546875	0xef = 1.11011_11 = -0b1.11*2 <sup>11</sup> = -3584
0x30 = 0.01100_00 = +0b1.00*2 <sup>-4</sup> = 0.0625	0x70 = 0.11100_00 = +0b1.00*2 <sup>12</sup> = 4096	0xb0 = 1.01100_00 = -0b1.00*2 <sup>-4</sup> = -0.0625	0xf0 = 1.11100_00 = -0b1.00*2 <sup>12</sup> = -4096
0x31 = 0.01100_01 = +0b1.01*2 <sup>-4</sup> = 0.078125	0x71 = 0.11100_01 = +0b1.01*2 <sup>12</sup> = 5120	0xb1 = 1.01100_01 = -0b1.01*2 <sup>-4</sup> = -0.078125	0xf1 = 1.11100_01 = -0b1.01*2 <sup>12</sup> = -5120
0x32 = 0.01100_10 = +0b1.10*2 <sup>-4</sup> = 0.09375	0x72 = 0.11100_10 = +0b1.10*2 <sup>12</sup> = 6144	0xb2 = 1.01100_10 = -0b1.10*2 <sup>-4</sup> = -0.09375	0xf2 = 1.11100_10 = -0b1.10*2 <sup>12</sup> = -6144
0x33 = 0.01100_11 = +0b1.11*2 <sup>-4</sup> = 0.109375	0x73 = 0.11100_11 = +0b1.11*2 <sup>12</sup> = 7168	0xb3 = 1.01100_11 = -0b1.11*2 <sup>-4</sup> = -0.109375	0xf3 = 1.11100_11 = -0b1.11*2 <sup>12</sup> = -7168
0x34 = 0.01101_00 = +0b1.00*2 <sup>-3</sup> = 0.125	0x74 = 0.11101_00 = +0b1.00*2 <sup>13</sup> = 8192	0xb4 = 1.01101_00 = -0b1.00*2 <sup>-3</sup> = -0.125	0xf4 = 1.11101_00 = -0b1.00*2 <sup>13</sup> = -8192
0x35 = 0.01101_01 = +0b1.01*2 <sup>-3</sup> = 0.15625	0x75 = 0.11101_01 = +0b1.01*2 <sup>13</sup> = 10240	0xb5 = 1.01101_01 = -0b1.01*2 <sup>-3</sup> = -0.15625	0xf5 = 1.11101_01 = -0b1.01*2 <sup>13</sup> = -10240
0x36 = 0.01101_10 = +0b1.10*2 <sup>-3</sup> = 0.1875	0x76 = 0.11101_10 = +0b1.10*2 <sup>13</sup> = 12288	0xb6 = 1.01101_10 = -0b1.10*2 <sup>-3</sup> = -0.1875	0xf6 = 1.11101_10 = -0b1.10*2 <sup>13</sup> = -12288
0x37 = 0.01101_11 = +0b1.11*2 <sup>-3</sup> = 0.21875	0x77 = 0.11101_11 = +0b1.11*2 <sup>13</sup> = 14336	0xb7 = 1.01101_11 = -0b1.11*2 <sup>-3</sup> = -0.21875	0xf7 = 1.11101_11 = -0b1.11*2 <sup>13</sup> = -14336
0x38 = 0.01110_00 = +0b1.00*2 <sup>-2</sup> = 0.25	0x78 = 0.11110_00 = +0b1.00*2 <sup>14</sup> = 16384	0xb8 = 1.01110_00 = -0b1.00*2 <sup>-2</sup> = -0.25	0xf8 = 1.11110_00 = -0b1.00*2 <sup>14</sup> = -16384
0x39 = 0.01110_01 = +0b1.01*2 <sup>-2</sup> = 0.3125	0x79 = 0.11110_01 = +0b1.01*2 <sup>14</sup> = 20480	0xb9 = 1.01110_01 = -0b1.01*2 <sup>-2</sup> = -0.3125	0xf9 = 1.11110_01 = -0b1.01*2 <sup>14</sup> = -20480
0x3a = 0.01110_10 = +0b1.10*2 <sup>-2</sup> = 0.375	0x7a = 0.11110_10 = +0b1.10*2 <sup>14</sup> = 24576	0xba = 1.01110_10 = -0b1.10*2 <sup>-2</sup> = -0.375	0xfa = 1.11110_10 = -0b1.10*2 <sup>14</sup> = -24576
0x3b = 0.01110_11 = +0b1.11*2 <sup>-2</sup> = 0.4375	0x7b = 0.11110_11 = +0b1.11*2 <sup>14</sup> = 28672	0xbb = 1.01110_11 = -0b1.11*2 <sup>-2</sup> = -0.4375	0xfb = 1.11110_11 = -0b1.11*2 <sup>14</sup> = -28672
0x3c = 0.01111_00 = +0b1.00*2 <sup>-1</sup> = 0.5	0x7c = 0.11111_00 = +0b1.00*2 <sup>15</sup> = 32768	0xbc = 1.01111_00 = -0b1.00*2 <sup>-1</sup> = -0.5	0xfc = 1.11111_00 = -0b1.00*2 <sup>15</sup> = -32768
0x3d = 0.01111_01 = +0b1.01*2 <sup>-1</sup> = 0.625	0x7d = 0.11111_01 = +0b1.01*2 <sup>15</sup> = 40960	0xbd = 1.01111_01 = -0b1.01*2 <sup>-1</sup> = -0.625	0xfd = 1.11111_01 = -0b1.01*2 <sup>15</sup> = -40960
0x3e = 0.01111_10 = +0b1.10*2 <sup>-1</sup> = 0.75	0x7e = 0.11111_10 = +0b1.10*2 <sup>15</sup> = 49152	0xbe = 1.01111_10 = -0b1.10*2 <sup>-1</sup> = -0.75	0xfe = 1.11111_10 = -0b1.10*2 <sup>15</sup> = -49152
0x3f = 0.01111_11 = +0b1.11*2 <sup>-1</sup> = 0.875	0x7f = 0.11111_11 = +Inf	0xbf = 1.01111_11 = -0b1.11*2 <sup>-1</sup> = -0.875	0xff = 1.11111_11 = -Inf

# Value table: binary8p4

0x00 = 0 0000 000 = +0b0.000*2 <sup>-7</sup> = 0	0x40 = 0 1000 000 = +0b1.000*2 <sup>0</sup> = 1	0x80 = 1 0000 000 = NaN	0xc0 = 1 1000 000 = -0b1.000*2 <sup>0</sup> = -1
0x01 = 0 0000 001 = +0b0.001*2 <sup>-7</sup> = 0.000976562	0x41 = 0 1000 001 = +0b1.001*2 <sup>0</sup> = 1.125	0x81 = 1 0000 001 = -0b0.001*2 <sup>-7</sup> = -0.000976562	0xc1 = 1 1000 001 = -0b1.001*2 <sup>0</sup> = -1.125
0x02 = 0 0000 010 = +0b0.010*2 <sup>-7</sup> = 0.00195312	0x42 = 0 1000 010 = +0b1.010*2 <sup>0</sup> = 1.25	0x82 = 1 0000 010 = -0b0.010*2 <sup>-7</sup> = -0.00195312	0xc2 = 1 1000 010 = -0b1.010*2 <sup>0</sup> = -1.25
0x03 = 0 0000 011 = +0b0.011*2 <sup>-7</sup> = 0.00292969	0x43 = 0 1000 011 = +0b1.011*2 <sup>0</sup> = 1.375	0x83 = 1 0000 011 = -0b0.011*2 <sup>-7</sup> = -0.00292969	0xc3 = 1 1000 011 = -0b1.011*2 <sup>0</sup> = -1.375
0x04 = 0 0000 100 = +0b0.100*2 <sup>-7</sup> = 0.00390625	0x44 = 0 1000 100 = +0b1.100*2 <sup>0</sup> = 1.5	0x84 = 1 0000 100 = -0b0.100*2 <sup>-7</sup> = -0.00390625	0xc4 = 1 1000 100 = -0b1.100*2 <sup>0</sup> = -1.5
0x05 = 0 0000 101 = +0b0.101*2 <sup>-7</sup> = 0.00488281	0x45 = 0 1000 101 = +0b1.101*2 <sup>0</sup> = 1.625	0x85 = 1 0000 101 = -0b0.101*2 <sup>-7</sup> = -0.00488281	0xc5 = 1 1000 101 = -0b1.101*2 <sup>0</sup> = -1.625
0x06 = 0 0000 110 = +0b0.110*2 <sup>-7</sup> = 0.00585938	0x46 = 0 1000 110 = +0b1.110*2 <sup>0</sup> = 1.75	0x86 = 1 0000 110 = -0b0.110*2 <sup>-7</sup> = -0.00585938	0xc6 = 1 1000 110 = -0b1.110*2 <sup>0</sup> = -1.75
0x07 = 0 0000 111 = +0b0.111*2 <sup>-7</sup> = 0.00683594	0x47 = 0 1000 111 = +0b1.111*2 <sup>0</sup> = 1.875	0x87 = 1 0000 111 = -0b0.111*2 <sup>-7</sup> = -0.00683594	0xc7 = 1 1000 111 = -0b1.111*2 <sup>0</sup> = -1.875
0x08 = 0 0001 000 = +0b1.000*2 <sup>-6</sup> = 0.0078125	0x48 = 0 1001 000 = +0b1.000*2 <sup>-1</sup> = 2	0x88 = 1 0001 000 = -0b1.000*2 <sup>-6</sup> = -0.0078125	0xc8 = 1 1001 000 = -0b1.000*2 <sup>-1</sup> = -2
0x09 = 0 0001 001 = +0b1.001*2 <sup>-6</sup> = 0.00878906	0x49 = 0 1001 001 = +0b1.001*2 <sup>-1</sup> = 2.25	0x89 = 1 0001 001 = -0b1.001*2 <sup>-6</sup> = -0.00878906	0xc9 = 1 1001 001 = -0b1.001*2 <sup>-1</sup> = -2.25
0x0a = 0 0001 010 = +0b1.010*2 <sup>-6</sup> = 0.00976562	0x4a = 0 1001 010 = +0b1.010*2 <sup>-1</sup> = 2.5	0x8a = 1 0001 010 = -0b1.010*2 <sup>-6</sup> = -0.00976562	0xca = 1 1001 010 = -0b1.010*2 <sup>-1</sup> = -2.5
0x0b = 0 0001 011 = +0b1.011*2 <sup>-6</sup> = 0.0107422	0x4b = 0 1001 011 = +0b1.011*2 <sup>-1</sup> = 2.75	0x8b = 1 0001 011 = -0b1.011*2 <sup>-6</sup> = -0.0107422	0xcb = 1 1001 011 = -0b1.011*2 <sup>-1</sup> = -2.75
0x0c = 0 0001 100 = +0b1.100*2 <sup>-6</sup> = 0.0117188	0x4c = 0 1001 100 = +0b1.100*2 <sup>-1</sup> = 3	0x8c = 1 0001 100 = -0b1.100*2 <sup>-6</sup> = -0.0117188	0xcc = 1 1001 100 = -0b1.100*2 <sup>-1</sup> = -3
0x0d = 0 0001 101 = +0b1.101*2 <sup>-6</sup> = 0.0126953	0x4d = 0 1001 101 = +0b1.101*2 <sup>-1</sup> = 3.25	0x8d = 1 0001 101 = -0b1.101*2 <sup>-6</sup> = -0.0126953	0xcd = 1 1001 101 = -0b1.101*2 <sup>-1</sup> = -3.25
0x0e = 0 0001 110 = +0b1.110*2 <sup>-6</sup> = 0.0136719	0x4e = 0 1001 110 = +0b1.110*2 <sup>-1</sup> = 3.5	0x8e = 1 0001 110 = -0b1.110*2 <sup>-6</sup> = -0.0136719	0xce = 1 1001 110 = -0b1.110*2 <sup>-1</sup> = -3.5
0x0f = 0 0001 111 = +0b1.111*2 <sup>-6</sup> = 0.0146484	0x4f = 0 1001 111 = +0b1.111*2 <sup>-1</sup> = 3.75	0x8f = 1 0001 111 = -0b1.111*2 <sup>-6</sup> = -0.0146484	0xcf = 1 1001 111 = -0b1.111*2 <sup>-1</sup> = -3.75
0x10 = 0 0010 000 = +0b1.000*2 <sup>-5</sup> = 0.015625	0x50 = 0 1010 000 = +0b1.000*2 <sup>-2</sup> = 4	0x90 = 1 0010 000 = -0b1.000*2 <sup>-5</sup> = -0.015625	0xd0 = 1 1010 000 = -0b1.000*2 <sup>-2</sup> = -4
0x11 = 0 0010 001 = +0b1.001*2 <sup>-5</sup> = 0.0175781	0x51 = 0 1010 001 = +0b1.001*2 <sup>-2</sup> = 4.5	0x91 = 1 0010 001 = -0b1.001*2 <sup>-5</sup> = -0.0175781	0xd1 = 1 1010 001 = -0b1.001*2 <sup>-2</sup> = -4.5
0x12 = 0 0010 010 = +0b1.010*2 <sup>-5</sup> = 0.0195312	0x52 = 0 1010 010 = +0b1.010*2 <sup>-2</sup> = 5	0x92 = 1 0010 010 = -0b1.010*2 <sup>-5</sup> = -0.0195312	0xd2 = 1 1010 010 = -0b1.010*2 <sup>-2</sup> = -5
0x13 = 0 0010 011 = +0b1.011*2 <sup>-5</sup> = 0.0214844	0x53 = 0 1010 011 = +0b1.011*2 <sup>-2</sup> = 5.5	0x93 = 1 0010 011 = -0b1.011*2 <sup>-5</sup> = -0.0214844	0xd3 = 1 1010 011 = -0b1.011*2 <sup>-2</sup> = -5.5
0x14 = 0 0010 100 = +0b1.100*2 <sup>-5</sup> = 0.0234375	0x54 = 0 1010 100 = +0b1.100*2 <sup>-2</sup> = 6	0x94 = 1 0010 100 = -0b1.100*2 <sup>-5</sup> = -0.0234375	0xd4 = 1 1010 100 = -0b1.100*2 <sup>-2</sup> = -6
0x15 = 0 0010 101 = +0b1.101*2 <sup>-5</sup> = 0.0253906	0x55 = 0 1010 101 = +0b1.101*2 <sup>-2</sup> = 6.5	0x95 = 1 0010 101 = -0b1.101*2 <sup>-5</sup> = -0.0253906	0xd5 = 1 1010 101 = -0b1.101*2 <sup>-2</sup> = -6.5
0x16 = 0 0010 110 = +0b1.110*2 <sup>-5</sup> = 0.0273438	0x56 = 0 1010 110 = +0b1.110*2 <sup>-2</sup> = 7	0x96 = 1 0010 110 = -0b1.110*2 <sup>-5</sup> = -0.0273438	0xd6 = 1 1010 110 = -0b1.110*2 <sup>-2</sup> = -7
0x17 = 0 0010 111 = +0b1.111*2 <sup>-5</sup> = 0.0292969	0x57 = 0 1010 111 = +0b1.111*2 <sup>-2</sup> = 7.5	0x97 = 1 0010 111 = -0b1.111*2 <sup>-5</sup> = -0.0292969	0xd7 = 1 1010 111 = -0b1.111*2 <sup>-2</sup> = -7.5
0x18 = 0 0011 000 = +0b1.000*2 <sup>-4</sup> = 0.03125	0x58 = 0 1011 000 = +0b1.000*2 <sup>-3</sup> = 8	0x98 = 1 0011 000 = -0b1.000*2 <sup>-4</sup> = -0.03125	0xd8 = 1 1011 000 = -0b1.000*2 <sup>-3</sup> = -8
0x19 = 0 0011 001 = +0b1.001*2 <sup>-4</sup> = 0.0351562	0x59 = 0 1011 001 = +0b1.001*2 <sup>-3</sup> = 9	0x99 = 1 0011 001 = -0b1.001*2 <sup>-4</sup> = -0.0351562	0xd9 = 1 1011 001 = -0b1.001*2 <sup>-3</sup> = -9
0x1a = 0 0011 010 = +0b1.010*2 <sup>-4</sup> = 0.0390625	0x5a = 0 1011 010 = +0b1.010*2 <sup>-3</sup> = 10	0x9a = 1 0011 010 = -0b1.010*2 <sup>-4</sup> = -0.0390625	0xda = 1 1011 010 = -0b1.010*2 <sup>-3</sup> = -10
0x1b = 0 0011 011 = +0b1.011*2 <sup>-4</sup> = 0.0429688	0x5b = 0 1011 011 = +0b1.011*2 <sup>-3</sup> = 11	0x9b = 1 0011 011 = -0b1.011*2 <sup>-4</sup> = -0.0429688	0xdb = 1 1011 011 = -0b1.011*2 <sup>-3</sup> = -11
0x1c = 0 0011 100 = +0b1.100*2 <sup>-4</sup> = 0.046875	0x5c = 0 1011 100 = +0b1.100*2 <sup>-3</sup> = 12	0x9c = 1 0011 100 = -0b1.100*2 <sup>-4</sup> = -0.046875	0xdc = 1 1011 100 = -0b1.100*2 <sup>-3</sup> = -12
0x1d = 0 0011 101 = +0b1.101*2 <sup>-4</sup> = 0.0507812	0x5d = 0 1011 101 = +0b1.101*2 <sup>-3</sup> = 13	0x9d = 1 0011 101 = -0b1.101*2 <sup>-4</sup> = -0.0507812	0xdd = 1 1011 101 = -0b1.101*2 <sup>-3</sup> = -13
0x1e = 0 0011 110 = +0b1.110*2 <sup>-4</sup> = 0.0546875	0x5e = 0 1011 110 = +0b1.110*2 <sup>-3</sup> = 14	0x9e = 1 0011 110 = -0b1.110*2 <sup>-4</sup> = -0.0546875	0xde = 1 1011 110 = -0b1.110*2 <sup>-3</sup> = -14
0x1f = 0 0011 111 = +0b1.111*2 <sup>-4</sup> = 0.0585938	0x5f = 0 1011 111 = +0b1.111*2 <sup>-3</sup> = 15	0x9f = 1 0011 111 = -0b1.111*2 <sup>-4</sup> = -0.0585938	0xdf = 1 1011 111 = -0b1.111*2 <sup>-3</sup> = -15
0x20 = 0 0100 000 = +0b1.000*2 <sup>-3</sup> = 0.0625	0x60 = 0 1100 000 = +0b1.000*2 <sup>-2</sup> = 16	0xa0 = 1 0100 000 = -0b1.000*2 <sup>-3</sup> = -0.0625	0xe0 = 1 1100 000 = -0b1.000*2 <sup>-2</sup> = -16
0x21 = 0 0100 001 = +0b1.001*2 <sup>-3</sup> = 0.0703125	0x61 = 0 1100 001 = +0b1.001*2 <sup>-2</sup> = 18	0xa1 = 1 0100 001 = -0b1.001*2 <sup>-3</sup> = -0.0703125	0xe1 = 1 1100 001 = -0b1.001*2 <sup>-2</sup> = -18
0x22 = 0 0100 010 = +0b1.010*2 <sup>-3</sup> = 0.078125	0x62 = 0 1100 010 = +0b1.010*2 <sup>-2</sup> = 20	0xa2 = 1 0100 010 = -0b1.010*2 <sup>-3</sup> = -0.078125	0xe2 = 1 1100 010 = -0b1.010*2 <sup>-2</sup> = -20
0x23 = 0 0100 011 = +0b1.011*2 <sup>-3</sup> = 0.0859375	0x63 = 0 1100 011 = +0b1.011*2 <sup>-2</sup> = 22	0xa3 = 1 0100 011 = -0b1.011*2 <sup>-3</sup> = -0.0859375	0xe3 = 1 1100 011 = -0b1.011*2 <sup>-2</sup> = -22
0x24 = 0 0100 100 = +0b1.100*2 <sup>-3</sup> = 0.09375	0x64 = 0 1100 100 = +0b1.100*2 <sup>-2</sup> = 24	0xa4 = 1 0100 100 = -0b1.100*2 <sup>-3</sup> = -0.09375	0xe4 = 1 1100 100 = -0b1.100*2 <sup>-2</sup> = -24
0x25 = 0 0100 101 = +0b1.101*2 <sup>-3</sup> = 0.101562	0x65 = 0 1100 101 = +0b1.101*2 <sup>-2</sup> = 26	0xa5 = 1 0100 101 = -0b1.101*2 <sup>-3</sup> = -0.101562	0xe5 = 1 1100 101 = -0b1.101*2 <sup>-2</sup> = -26
0x26 = 0 0100 110 = +0b1.110*2 <sup>-3</sup> = 0.109375	0x66 = 0 1100 110 = +0b1.110*2 <sup>-2</sup> = 28	0xa6 = 1 0100 110 = -0b1.110*2 <sup>-3</sup> = -0.109375	0xe6 = 1 1100 110 = -0b1.110*2 <sup>-2</sup> = -28
0x27 = 0 0100 111 = +0b1.111*2 <sup>-3</sup> = 0.117188	0x67 = 0 1100 111 = +0b1.111*2 <sup>-2</sup> = 30	0xa7 = 1 0100 111 = -0b1.111*2 <sup>-3</sup> = -0.117188	0xe7 = 1 1100 111 = -0b1.111*2 <sup>-2</sup> = -30
0x28 = 0 0101 000 = +0b1.000*2 <sup>-2</sup> = 0.125	0x68 = 0 1101 000 = +0b1.000*2 <sup>-1</sup> = 32	0xa8 = 1 0101 000 = -0b1.000*2 <sup>-2</sup> = -0.125	0xe8 = 1 1101 000 = -0b1.000*2 <sup>-1</sup> = -32
0x29 = 0 0101 001 = +0b1.001*2 <sup>-2</sup> = 0.140625	0x69 = 0 1101 001 = +0b1.001*2 <sup>-1</sup> = 36	0xa9 = 1 0101 001 = -0b1.001*2 <sup>-2</sup> = -0.140625	0xe9 = 1 1101 001 = -0b1.001*2 <sup>-1</sup> = -36
0x2a = 0 0101 010 = +0b1.010*2 <sup>-2</sup> = 0.15625	0x6a = 0 1101 010 = +0b1.010*2 <sup>-1</sup> = 40	0xaa = 1 0101 010 = -0b1.010*2 <sup>-2</sup> = -0.15625	0xea = 1 1101 010 = -0b1.010*2 <sup>-1</sup> = -40
0x2b = 0 0101 011 = +0b1.011*2 <sup>-2</sup> = 0.171875	0x6b = 0 1101 011 = +0b1.011*2 <sup>-1</sup> = 44	0xab = 1 0101 011 = -0b1.011*2 <sup>-2</sup> = -0.171875	0xeb = 1 1101 011 = -0b1.011*2 <sup>-1</sup> = -44
0x2c = 0 0101 100 = +0b1.100*2 <sup>-2</sup> = 0.1875	0x6c = 0 1101 100 = +0b1.100*2 <sup>-1</sup> = 48	0xac = 1 0101 100 = -0b1.100*2 <sup>-2</sup> = -0.1875	0xec = 1 1101 100 = -0b1.100*2 <sup>-1</sup> = -48
0x2d = 0 0101 101 = +0b1.101*2 <sup>-2</sup> = 0.203125	0x6d = 0 1101 101 = +0b1.101*2 <sup>-1</sup> = 52	0xad = 1 0101 101 = -0b1.101*2 <sup>-2</sup> = -0.203125	0xed = 1 1101 101 = -0b1.101*2 <sup>-1</sup> = -52
0x2e = 0 0101 110 = +0b1.110*2 <sup>-2</sup> = 0.21875	0x6e = 0 1101 110 = +0b1.110*2 <sup>-1</sup> = 56	0xae = 1 0101 110 = -0b1.110*2 <sup>-2</sup> = -0.21875	0xee = 1 1101 110 = -0b1.110*2 <sup>-1</sup> = -56
0x2f = 0 0101 111 = +0b1.111*2 <sup>-2</sup> = 0.234375	0x6f = 0 1101 111 = +0b1.111*2 <sup>-1</sup> = 60	0xaf = 1 0101 111 = -0b1.111*2 <sup>-2</sup> = -0.234375	0xef = 1 1101 111 = -0b1.111*2 <sup>-1</sup> = -60
0x30 = 0 0110 000 = +0b1.000*2 <sup>-1</sup> = 0.25	0x70 = 0 1110 000 = +0b1.000*2 <sup>-0</sup> = 64	0xb0 = 1 0110 000 = -0b1.000*2 <sup>-1</sup> = -0.25	0xf0 = 1 1110 000 = -0b1.000*2 <sup>-0</sup> = -64
0x31 = 0 0110 001 = +0b1.001*2 <sup>-1</sup> = 0.28125	0x71 = 0 1110 001 = +0b1.001*2 <sup>-0</sup> = 72	0xb1 = 1 0110 001 = -0b1.001*2 <sup>-1</sup> = -0.28125	0xf1 = 1 1110 001 = -0b1.001*2 <sup>-0</sup> = -72
0x32 = 0 0110 010 = +0b1.010*2 <sup>-1</sup> = 0.3125	0x72 = 0 1110 010 = +0b1.010*2 <sup>-0</sup> = 80	0xb2 = 1 0110 010 = -0b1.010*2 <sup>-1</sup> = -0.3125	0xf2 = 1 1110 010 = -0b1.010*2 <sup>-0</sup> = -80
0x33 = 0 0110 011 = +0b1.011*2 <sup>-1</sup> = 0.34375	0x73 = 0 1110 011 = +0b1.011*2 <sup>-0</sup> = 88	0xb3 = 1 0110 011 = -0b1.011*2 <sup>-1</sup> = -0.34375	0xf3 = 1 1110 011 = -0b1.011*2 <sup>-0</sup> = -88
0x34 = 0 0110 100 = +0b1.100*2 <sup>-1</sup> = 0.375	0x74 = 0 1110 100 = +0b1.100*2 <sup>-0</sup> = 96	0xb4 = 1 0110 100 = -0b1.100*2 <sup>-1</sup> = -0.375	0xf4 = 1 1110 100 = -0b1.100*2 <sup>-0</sup> = -96
0x35 = 0 0110 101 = +0b1.101*2 <sup>-1</sup> = 0.40625	0x75 = 0 1110 101 = +0b1.101*2 <sup>-0</sup> = 104	0xb5 = 1 0110 101 = -0b1.101*2 <sup>-1</sup> = -0.40625	0xf5 = 1 1110 101 = -0b1.101*2 <sup>-0</sup> = -104
0x36 = 0 0110 110 = +0b1.110*2 <sup>-1</sup> = 0.4375	0x76 = 0 1110 110 = +0b1.110*2 <sup>-0</sup> = 112	0xb6 = 1 0110 110 = -0b1.110*2 <sup>-1</sup> = -0.4375	0xf6 = 1 1110 110 = -0b1.110*2 <sup>-0</sup> = -112
0x37 = 0 0110 111 = +0b1.111*2 <sup>-1</sup> = 0.46875	0x77 = 0 1110 111 = +0b1.111*2 <sup>-0</sup> = 120	0xb7 = 1 0110 111 = -0b1.111*2 <sup>-1</sup> = -0.46875	0xf7 = 1 1110 111 = -0b1.111*2 <sup>-0</sup> = -120
0x38 = 0 0111 000 = +0b1.000*2 <sup>-0</sup> = 0.5	0x78 = 0 1111 000 = +0b1.000*2 <sup>-0</sup> = 128	0xb8 = 1 0111 000 = -0b1.000*2 <sup>-0</sup> = -0.5	0xf8 = 1 1111 000 = -0b1.000*2 <sup>-0</sup> = -128
0x39 = 0 0111 001 = +0b1.001*2 <sup>-0</sup> = 0.5625	0x79 = 0 1111 001 = +0b1.001*2 <sup>-0</sup> = 144	0xb9 = 1 0111 001 = -0b1.001*2 <sup>-0</sup> = -0.5625	0xf9 = 1 1111 001 = -0b1.001*2 <sup>-0</sup> = -144
0x3a = 0 0111 010 = +0b1.010*2 <sup>-0</sup> = 0.625	0x7a = 0 1111 010 = +0b1.010*2 <sup>-0</sup> = 160	0xba = 1 0111 010 = -0b1.010*2 <sup>-0</sup> = -0.625	0xfa = 1 1111 010 = -0b1.010*2 <sup>-0</sup> = -160
0x3b = 0 0111 011 = +0b1.011*2 <sup>-0</sup> = 0.6875	0x7b = 0 1111 011 = +0b1.011*2 <sup>-0</sup> = 176	0xbb = 1 0111 011 = -0b1.011*2 <sup>-0</sup> = -0.6875	0xfb = 1 1111 011 = -0b1.011*2 <sup>-0</sup> = -176
0x3c = 0 0111 100 = +0b1.100*2 <sup>-0</sup> = 0.75	0x7c = 0 1111 100 = +0b1.100*2 <sup>-0</sup> = 192	0xbc = 1 0111 100 = -0b1.100*2 <sup>-0</sup> = -0.75	0xfc = 1 1111 100 = -0b1.100*2 <sup>-0</sup> = -192
0x3d = 0 0111 101 = +0b1.101*2 <sup>-0</sup> = 0.8125	0x7d = 0 1111 101 = +0b1.101*2 <sup>-0</sup> = 208	0xbd = 1 0111 101 = -0b1.101*2 <sup>-0</sup> = -0.8125	0xfd = 1 1111 101 = -0b1.101*2 <sup>-0</sup> = -208
0x3e = 0 0111 110 = +0b1.110*2 <sup>-0</sup> = 0.875	0x7e = 0 1111 110 = +0b1.110*2 <sup>-0</sup> = 224	0xbe = 1 0111 110 = -0b1.110*2 <sup>-0</sup> = -0.875	0xfe = 1 1111 110 = -0b1.110*2 <sup>-0</sup> = -224
0x3f = 0 0111 111 = +0b1.111*2 <sup>-0</sup> = 0.9375	0x7f = 0 1111 111 = +Inf	0xbf = 1 0111 111 = -0b1.111*2 <sup>-0</sup> = -0.9375	0xff = 1 1111 111 = -Inf

# Value table: binary8p5

0x00 = 0.000.0000 = +0b0.0000*2 <sup>-3</sup> = 0	0x40 = 0.100.0000 = +0b1.0000*2 <sup>0</sup> = 1	0x80 = 1.000.0000 = NaN	0xc0 = 1.100.0000 = -0b1.0000*2 <sup>0</sup> = -1
0x01 = 0.000.0001 = +0b0.0001*2 <sup>-3</sup> = 0.0078125	0x41 = 0.100.0001 = +0b1.0001*2 <sup>0</sup> = 1.0625	0x81 = 1.000.0001 = -0b0.0001*2 <sup>-3</sup> = -0.0078125	0xc1 = 1.100.0001 = -0b1.0001*2 <sup>0</sup> = -1.0625
0x02 = 0.000.0010 = +0b0.0010*2 <sup>-3</sup> = 0.015625	0x42 = 0.100.0010 = +0b1.0010*2 <sup>0</sup> = 1.125	0x82 = 1.000.0010 = -0b0.0010*2 <sup>-3</sup> = -0.015625	0xc2 = 1.100.0010 = -0b1.0010*2 <sup>0</sup> = -1.125
0x03 = 0.000.0011 = +0b0.0011*2 <sup>-3</sup> = 0.0234375	0x43 = 0.100.0011 = +0b1.0011*2 <sup>0</sup> = 1.1875	0x83 = 1.000.0011 = -0b0.0011*2 <sup>-3</sup> = -0.0234375	0xc3 = 1.100.0011 = -0b1.0011*2 <sup>0</sup> = -1.1875
0x04 = 0.000.0100 = +0b0.0100*2 <sup>-3</sup> = 0.03125	0x44 = 0.100.0100 = +0b1.0100*2 <sup>0</sup> = 1.25	0x84 = 1.000.0100 = -0b0.0100*2 <sup>-3</sup> = -0.03125	0xc4 = 1.100.0100 = -0b1.0100*2 <sup>0</sup> = -1.25
0x05 = 0.000.0101 = +0b0.0101*2 <sup>-3</sup> = 0.0390625	0x45 = 0.100.0101 = +0b1.0101*2 <sup>0</sup> = 1.3125	0x85 = 1.000.0101 = -0b0.0101*2 <sup>-3</sup> = -0.0390625	0xc5 = 1.100.0101 = -0b1.0101*2 <sup>0</sup> = -1.3125
0x06 = 0.000.0110 = +0b0.0110*2 <sup>-3</sup> = 0.046875	0x46 = 0.100.0110 = +0b1.0110*2 <sup>0</sup> = 1.375	0x86 = 1.000.0110 = -0b0.0110*2 <sup>-3</sup> = -0.046875	0xc6 = 1.100.0110 = -0b1.0110*2 <sup>0</sup> = -1.375
0x07 = 0.000.0111 = +0b0.0111*2 <sup>-3</sup> = 0.0546875	0x47 = 0.100.0111 = +0b1.0111*2 <sup>0</sup> = 1.4375	0x87 = 1.000.0111 = -0b0.0111*2 <sup>-3</sup> = -0.0546875	0xc7 = 1.100.0111 = -0b1.0111*2 <sup>0</sup> = -1.4375
0x08 = 0.000.1000 = +0b0.1000*2 <sup>-3</sup> = 0.0625	0x48 = 0.100.1000 = +0b1.1000*2 <sup>0</sup> = 1.5	0x88 = 1.000.1000 = -0b0.1000*2 <sup>-3</sup> = -0.0625	0xc8 = 1.100.1000 = -0b1.1000*2 <sup>0</sup> = -1.5
0x09 = 0.000.1001 = +0b0.1001*2 <sup>-3</sup> = 0.0703125	0x49 = 0.100.1001 = +0b1.1001*2 <sup>0</sup> = 1.5625	0x89 = 1.000.1001 = -0b0.1001*2 <sup>-3</sup> = -0.0703125	0xc9 = 1.100.1001 = -0b1.1001*2 <sup>0</sup> = -1.5625
0x0a = 0.000.1010 = +0b0.1010*2 <sup>-3</sup> = 0.078125	0x4a = 0.100.1010 = +0b1.1010*2 <sup>0</sup> = 1.625	0x8a = 1.000.1010 = -0b0.1010*2 <sup>-3</sup> = -0.078125	0xca = 1.100.1010 = -0b1.1010*2 <sup>0</sup> = -1.625
0x0b = 0.000.1011 = +0b0.1011*2 <sup>-3</sup> = 0.0859375	0x4b = 0.100.1011 = +0b1.1011*2 <sup>0</sup> = 1.6875	0x8b = 1.000.1011 = -0b0.1011*2 <sup>-3</sup> = -0.0859375	0xcb = 1.100.1011 = -0b1.1011*2 <sup>0</sup> = -1.6875
0x0c = 0.000.1100 = +0b0.1100*2 <sup>-3</sup> = 0.09375	0x4c = 0.100.1100 = +0b1.1100*2 <sup>0</sup> = 1.75	0x8c = 1.000.1100 = -0b0.1100*2 <sup>-3</sup> = -0.09375	0xcc = 1.100.1100 = -0b1.1100*2 <sup>0</sup> = -1.75
0x0d = 0.000.1101 = +0b0.1101*2 <sup>-3</sup> = 0.101562	0x4d = 0.100.1101 = +0b1.1101*2 <sup>0</sup> = 1.8125	0x8d = 1.000.1101 = -0b0.1101*2 <sup>-3</sup> = -0.101562	0xcd = 1.100.1101 = -0b1.1101*2 <sup>0</sup> = -1.8125
0x0e = 0.000.1110 = +0b0.1110*2 <sup>-3</sup> = 0.109375	0x4e = 0.100.1110 = +0b1.1110*2 <sup>0</sup> = 1.875	0x8e = 1.000.1110 = -0b0.1110*2 <sup>-3</sup> = -0.109375	0xce = 1.100.1110 = -0b1.1110*2 <sup>0</sup> = -1.875
0x0f = 0.000.1111 = +0b0.1111*2 <sup>-3</sup> = 0.117188	0x4f = 0.100.1111 = +0b1.1111*2 <sup>0</sup> = 1.9375	0x8f = 1.000.1111 = -0b0.1111*2 <sup>-3</sup> = -0.117188	0xcf = 1.100.1111 = -0b1.1111*2 <sup>0</sup> = -1.9375
0x10 = 0.001.0000 = +0b1.0000*2 <sup>-3</sup> = 0.125	0x50 = 0.101.0000 = +0b1.0000*2 <sup>-1</sup> = 2	0x90 = 1.001.0000 = -0b1.0000*2 <sup>-3</sup> = -0.125	0xd0 = 1.101.0000 = -0b1.0000*2 <sup>-1</sup> = -2
0x11 = 0.001.0001 = +0b1.0001*2 <sup>-3</sup> = 0.132812	0x51 = 0.101.0001 = +0b1.0001*2 <sup>-1</sup> = 2.125	0x91 = 1.001.0001 = -0b1.0001*2 <sup>-3</sup> = -0.132812	0xd1 = 1.101.0001 = -0b1.0001*2 <sup>-1</sup> = -2.125
0x12 = 0.001.0010 = +0b1.0010*2 <sup>-3</sup> = 0.140625	0x52 = 0.101.0010 = +0b1.0010*2 <sup>-1</sup> = 2.25	0x92 = 1.001.0010 = -0b1.0010*2 <sup>-3</sup> = -0.140625	0xd2 = 1.101.0010 = -0b1.0010*2 <sup>-1</sup> = -2.25
0x13 = 0.001.0011 = +0b1.0011*2 <sup>-3</sup> = 0.148438	0x53 = 0.101.0011 = +0b1.0011*2 <sup>-1</sup> = 2.375	0x93 = 1.001.0011 = -0b1.0011*2 <sup>-3</sup> = -0.148438	0xd3 = 1.101.0011 = -0b1.0011*2 <sup>-1</sup> = -2.375
0x14 = 0.001.0100 = +0b1.0100*2 <sup>-3</sup> = 0.15625	0x54 = 0.101.0100 = +0b1.0100*2 <sup>-1</sup> = 2.5	0x94 = 1.001.0100 = -0b1.0100*2 <sup>-3</sup> = -0.15625	0xd4 = 1.101.0100 = -0b1.0100*2 <sup>-1</sup> = -2.5
0x15 = 0.001.0101 = +0b1.0101*2 <sup>-3</sup> = 0.164062	0x55 = 0.101.0101 = +0b1.0101*2 <sup>-1</sup> = 2.625	0x95 = 1.001.0101 = -0b1.0101*2 <sup>-3</sup> = -0.164062	0xd5 = 1.101.0101 = -0b1.0101*2 <sup>-1</sup> = -2.625
0x16 = 0.001.0110 = +0b1.0110*2 <sup>-3</sup> = 0.171875	0x56 = 0.101.0110 = +0b1.0110*2 <sup>-1</sup> = 2.75	0x96 = 1.001.0110 = -0b1.0110*2 <sup>-3</sup> = -0.171875	0xd6 = 1.101.0110 = -0b1.0110*2 <sup>-1</sup> = -2.75
0x17 = 0.001.0111 = +0b1.0111*2 <sup>-3</sup> = 0.179688	0x57 = 0.101.0111 = +0b1.0111*2 <sup>-1</sup> = 2.875	0x97 = 1.001.0111 = -0b1.0111*2 <sup>-3</sup> = -0.179688	0xd7 = 1.101.0111 = -0b1.0111*2 <sup>-1</sup> = -2.875
0x18 = 0.001.1000 = +0b1.1000*2 <sup>-3</sup> = 0.1875	0x58 = 0.101.1000 = +0b1.1000*2 <sup>-1</sup> = 3	0x98 = 1.001.1000 = -0b1.1000*2 <sup>-3</sup> = -0.1875	0xd8 = 1.101.1000 = -0b1.1000*2 <sup>-1</sup> = -3
0x19 = 0.001.1001 = +0b1.1001*2 <sup>-3</sup> = 0.195312	0x59 = 0.101.1001 = +0b1.1001*2 <sup>-1</sup> = 3.125	0x99 = 1.001.1001 = -0b1.1001*2 <sup>-3</sup> = -0.195312	0xd9 = 1.101.1001 = -0b1.1001*2 <sup>-1</sup> = -3.125
0x1a = 0.001.1010 = +0b1.1010*2 <sup>-3</sup> = 0.203125	0x5a = 0.101.1010 = +0b1.1010*2 <sup>-1</sup> = 3.25	0x9a = 1.001.1010 = -0b1.1010*2 <sup>-3</sup> = -0.203125	0xda = 1.101.1010 = -0b1.1010*2 <sup>-1</sup> = -3.25
0x1b = 0.001.1011 = +0b1.1011*2 <sup>-3</sup> = 0.210938	0x5b = 0.101.1011 = +0b1.1011*2 <sup>-1</sup> = 3.375	0x9b = 1.001.1011 = -0b1.1011*2 <sup>-3</sup> = -0.210938	0xdb = 1.101.1011 = -0b1.1011*2 <sup>-1</sup> = -3.375
0x1c = 0.001.1100 = +0b1.1100*2 <sup>-3</sup> = 0.21875	0x5c = 0.101.1100 = +0b1.1100*2 <sup>-1</sup> = 3.5	0x9c = 1.001.1100 = -0b1.1100*2 <sup>-3</sup> = -0.21875	0xdc = 1.101.1100 = -0b1.1100*2 <sup>-1</sup> = -3.5
0x1d = 0.001.1101 = +0b1.1101*2 <sup>-3</sup> = 0.226562	0x5d = 0.101.1101 = +0b1.1101*2 <sup>-1</sup> = 3.625	0x9d = 1.001.1101 = -0b1.1101*2 <sup>-3</sup> = -0.226562	0xdd = 1.101.1101 = -0b1.1101*2 <sup>-1</sup> = -3.625
0x1e = 0.001.1110 = +0b1.1110*2 <sup>-3</sup> = 0.234375	0x5e = 0.101.1110 = +0b1.1110*2 <sup>-1</sup> = 3.75	0x9e = 1.001.1110 = -0b1.1110*2 <sup>-3</sup> = -0.234375	0xde = 1.101.1110 = -0b1.1110*2 <sup>-1</sup> = -3.75
0x1f = 0.001.1111 = +0b1.1111*2 <sup>-3</sup> = 0.242188	0x5f = 0.101.1111 = +0b1.1111*2 <sup>-1</sup> = 3.875	0x9f = 1.001.1111 = -0b1.1111*2 <sup>-3</sup> = -0.242188	0xdf = 1.101.1111 = -0b1.1111*2 <sup>-1</sup> = -3.875
0x20 = 0.010.0000 = +0b1.0000*2 <sup>-2</sup> = 0.25	0x60 = 0.110.0000 = +0b1.0000*2 <sup>-2</sup> = 4	0xa0 = 1.010.0000 = -0b1.0000*2 <sup>-2</sup> = -0.25	0xe0 = 1.110.0000 = -0b1.0000*2 <sup>-2</sup> = -4
0x21 = 0.010.0001 = +0b1.0001*2 <sup>-2</sup> = 0.265625	0x61 = 0.110.0001 = +0b1.0001*2 <sup>-2</sup> = 4.25	0xa1 = 1.010.0001 = -0b1.0001*2 <sup>-2</sup> = -0.265625	0xe1 = 1.110.0001 = -0b1.0001*2 <sup>-2</sup> = -4.25
0x22 = 0.010.0010 = +0b1.0010*2 <sup>-2</sup> = 0.28125	0x62 = 0.110.0010 = +0b1.0010*2 <sup>-2</sup> = 4.5	0xa2 = 1.010.0010 = -0b1.0010*2 <sup>-2</sup> = -0.28125	0xe2 = 1.110.0010 = -0b1.0010*2 <sup>-2</sup> = -4.5
0x23 = 0.010.0011 = +0b1.0011*2 <sup>-2</sup> = 0.296875	0x63 = 0.110.0011 = +0b1.0011*2 <sup>-2</sup> = 4.75	0xa3 = 1.010.0011 = -0b1.0011*2 <sup>-2</sup> = -0.296875	0xe3 = 1.110.0011 = -0b1.0011*2 <sup>-2</sup> = -4.75
0x24 = 0.010.0100 = +0b1.0100*2 <sup>-2</sup> = 0.3125	0x64 = 0.110.0100 = +0b1.0100*2 <sup>-2</sup> = 5	0xa4 = 1.010.0100 = -0b1.0100*2 <sup>-2</sup> = -0.3125	0xe4 = 1.110.0100 = -0b1.0100*2 <sup>-2</sup> = -5
0x25 = 0.010.0101 = +0b1.0101*2 <sup>-2</sup> = 0.328125	0x65 = 0.110.0101 = +0b1.0101*2 <sup>-2</sup> = 5.25	0xa5 = 1.010.0101 = -0b1.0101*2 <sup>-2</sup> = -0.328125	0xe5 = 1.110.0101 = -0b1.0101*2 <sup>-2</sup> = -5.25
0x26 = 0.010.0110 = +0b1.0110*2 <sup>-2</sup> = 0.34375	0x66 = 0.110.0110 = +0b1.0110*2 <sup>-2</sup> = 5.5	0xa6 = 1.010.0110 = -0b1.0110*2 <sup>-2</sup> = -0.34375	0xe6 = 1.110.0110 = -0b1.0110*2 <sup>-2</sup> = -5.5
0x27 = 0.010.0111 = +0b1.0111*2 <sup>-2</sup> = 0.359375	0x67 = 0.110.0111 = +0b1.0111*2 <sup>-2</sup> = 5.75	0xa7 = 1.010.0111 = -0b1.0111*2 <sup>-2</sup> = -0.359375	0xe7 = 1.110.0111 = -0b1.0111*2 <sup>-2</sup> = -5.75
0x28 = 0.010.1000 = +0b1.1000*2 <sup>-2</sup> = 0.375	0x68 = 0.110.1000 = +0b1.1000*2 <sup>-2</sup> = 6	0xa8 = 1.010.1000 = -0b1.1000*2 <sup>-2</sup> = -0.375	0xe8 = 1.110.1000 = -0b1.1000*2 <sup>-2</sup> = -6
0x29 = 0.010.1001 = +0b1.1001*2 <sup>-2</sup> = 0.390625	0x69 = 0.110.1001 = +0b1.1001*2 <sup>-2</sup> = 6.25	0xa9 = 1.010.1001 = -0b1.1001*2 <sup>-2</sup> = -0.390625	0xe9 = 1.110.1001 = -0b1.1001*2 <sup>-2</sup> = -6.25
0x2a = 0.010.1010 = +0b1.1010*2 <sup>-2</sup> = 0.40625	0x6a = 0.110.1010 = +0b1.1010*2 <sup>-2</sup> = 6.5	0xaa = 1.010.1010 = -0b1.1010*2 <sup>-2</sup> = -0.40625	0xea = 1.110.1010 = -0b1.1010*2 <sup>-2</sup> = -6.5
0x2b = 0.010.1011 = +0b1.1011*2 <sup>-2</sup> = 0.421875	0x6b = 0.110.1011 = +0b1.1011*2 <sup>-2</sup> = 6.75	0xab = 1.010.1011 = -0b1.1011*2 <sup>-2</sup> = -0.421875	0xeb = 1.110.1011 = -0b1.1011*2 <sup>-2</sup> = -6.75
0x2c = 0.010.1100 = +0b1.1100*2 <sup>-2</sup> = 0.4375	0x6c = 0.110.1100 = +0b1.1100*2 <sup>-2</sup> = 7	0xac = 1.010.1100 = -0b1.1100*2 <sup>-2</sup> = -0.4375	0xec = 1.110.1100 = -0b1.1100*2 <sup>-2</sup> = -7
0x2d = 0.010.1101 = +0b1.1101*2 <sup>-2</sup> = 0.453125	0x6d = 0.110.1101 = +0b1.1101*2 <sup>-2</sup> = 7.25	0xad = 1.010.1101 = -0b1.1101*2 <sup>-2</sup> = -0.453125	0xed = 1.110.1101 = -0b1.1101*2 <sup>-2</sup> = -7.25
0x2e = 0.010.1110 = +0b1.1110*2 <sup>-2</sup> = 0.46875	0x6e = 0.110.1110 = +0b1.1110*2 <sup>-2</sup> = 7.5	0xae = 1.010.1110 = -0b1.1110*2 <sup>-2</sup> = -0.46875	0xee = 1.110.1110 = -0b1.1110*2 <sup>-2</sup> = -7.5
0x2f = 0.010.1111 = +0b1.1111*2 <sup>-2</sup> = 0.484375	0x6f = 0.110.1111 = +0b1.1111*2 <sup>-2</sup> = 7.75	0xaf = 1.010.1111 = -0b1.1111*2 <sup>-2</sup> = -0.484375	0xef = 1.110.1111 = -0b1.1111*2 <sup>-2</sup> = -7.75
0x30 = 0.011.0000 = +0b1.0000*2 <sup>-1</sup> = 0.5	0x70 = 0.111.0000 = +0b1.0000*2 <sup>-3</sup> = 8	0xb0 = 1.011.0000 = -0b1.0000*2 <sup>-1</sup> = -0.5	0xf0 = 1.111.0000 = -0b1.0000*2 <sup>-3</sup> = -8
0x31 = 0.011.0001 = +0b1.0001*2 <sup>-1</sup> = 0.53125	0x71 = 0.111.0001 = +0b1.0001*2 <sup>-3</sup> = 8.5	0xb1 = 1.011.0001 = -0b1.0001*2 <sup>-1</sup> = -0.53125	0xf1 = 1.111.0001 = -0b1.0001*2 <sup>-3</sup> = -8.5
0x32 = 0.011.0010 = +0b1.0010*2 <sup>-1</sup> = 0.5625	0x72 = 0.111.0010 = +0b1.0010*2 <sup>-3</sup> = 9	0xb2 = 1.011.0010 = -0b1.0010*2 <sup>-1</sup> = -0.5625	0xf2 = 1.111.0010 = -0b1.0010*2 <sup>-3</sup> = -9
0x33 = 0.011.0011 = +0b1.0011*2 <sup>-1</sup> = 0.59375	0x73 = 0.111.0011 = +0b1.0011*2 <sup>-3</sup> = 9.5	0xb3 = 1.011.0011 = -0b1.0011*2 <sup>-1</sup> = -0.59375	0xf3 = 1.111.0011 = -0b1.0011*2 <sup>-3</sup> = -9.5
0x34 = 0.011.0100 = +0b1.0100*2 <sup>-1</sup> = 0.625	0x74 = 0.111.0100 = +0b1.0100*2 <sup>-3</sup> = 10	0xb4 = 1.011.0100 = -0b1.0100*2 <sup>-1</sup> = -0.625	0xf4 = 1.111.0100 = -0b1.0100*2 <sup>-3</sup> = -10
0x35 = 0.011.0101 = +0b1.0101*2 <sup>-1</sup> = 0.65625	0x75 = 0.111.0101 = +0b1.0101*2 <sup>-3</sup> = 10.5	0xb5 = 1.011.0101 = -0b1.0101*2 <sup>-1</sup> = -0.65625	0xf5 = 1.111.0101 = -0b1.0101*2 <sup>-3</sup> = -10.5
0x36 = 0.011.0110 = +0b1.0110*2 <sup>-1</sup> = 0.6875	0x76 = 0.111.0110 = +0b1.0110*2 <sup>-3</sup> = 11	0xb6 = 1.011.0110 = -0b1.0110*2 <sup>-1</sup> = -0.6875	0xf6 = 1.111.0110 = -0b1.0110*2 <sup>-3</sup> = -11
0x37 = 0.011.0111 = +0b1.0111*2 <sup>-1</sup> = 0.71875	0x77 = 0.111.0111 = +0b1.0111*2 <sup>-3</sup> = 11.5	0xb7 = 1.011.0111 = -0b1.0111*2 <sup>-1</sup> = -0.71875	0xf7 = 1.111.0111 = -0b1.0111*2 <sup>-3</sup> = -11.5
0x38 = 0.011.1000 = +0b1.1000*2 <sup>-1</sup> = 0.75	0x78 = 0.111.1000 = +0b1.1000*2 <sup>-3</sup> = 12	0xb8 = 1.011.1000 = -0b1.1000*2 <sup>-1</sup> = -0.75	0xf8 = 1.111.1000 = -0b1.1000*2 <sup>-3</sup> = -12
0x39 = 0.011.1001 = +0b1.1001*2 <sup>-1</sup> = 0.78125	0x79 = 0.111.1001 = +0b1.1001*2 <sup>-3</sup> = 12.5	0xb9 = 1.011.1001 = -0b1.1001*2 <sup>-1</sup> = -0.78125	0xf9 = 1.111.1001 = -0b1.1001*2 <sup>-3</sup> = -12.5
0x3a = 0.011.1010 = +0b1.1010*2 <sup>-1</sup> = 0.8125	0x7a = 0.111.1010 = +0b1.1010*2 <sup>-3</sup> = 13	0xba = 1.011.1010 = -0b1.1010*2 <sup>-1</sup> = -0.8125	0xfa = 1.111.1010 = -0b1.1010*2 <sup>-3</sup> = -13
0x3b = 0.011.1011 = +0b1.1011*2 <sup>-1</sup> = 0.84375	0x7b = 0.111.1011 = +0b1.1011*2 <sup>-3</sup> = 13.5	0xbb = 1.011.1011 = -0b1.1011*2 <sup>-1</sup> = -0.84375	0xfb = 1.111.1011 = -0b1.1011*2 <sup>-3</sup> = -13.5
0x3c = 0.011.1100 = +0b1.1100*2 <sup>-1</sup> = 0.875	0x7c = 0.111.1100 = +0b1.1100*2 <sup>-3</sup> = 14	0xbc = 1.011.1100 = -0b1.1100*2 <sup>-1</sup> = -0.875	0xfc = 1.111.1100 = -0b1.1100*2 <sup>-3</sup> = -14
0x3d = 0.011.1101 = +0b1.1101*2 <sup>-1</sup> = 0.90625	0x7d = 0.111.1101 = +0b1.1101*2 <sup>-3</sup> = 14.5	0xbd = 1.011.1101 = -0b1.1101*2 <sup>-1</sup> = -0.90625	0xfd = 1.111.1101 = -0b1.1101*2 <sup>-3</sup> = -14.5
0x3e = 0.011.1110 = +0b1.1110*2 <sup>-1</sup> = 0.9375	0x7e = 0.111.1110 = +0b1.1110*2 <sup>-3</sup> = 15	0xbe = 1.011.1110 = -0b1.1110*2 <sup>-1</sup> = -0.9375	0xfe = 1.111.1110 = -0b1.1110*2 <sup>-3</sup> = -15
0x3f = 0.011.1111 = +0b1.1111*2 <sup>-1</sup> = 0.96875	0x7f = 0.111.1111 = +Inf	0xbf = 1.011.1111 = -0b1.1111*2 <sup>-1</sup> = -0.96875	0xff = 1.111.1