# 2.1 ASCII and Unicode

## Bits: 0's and 1's

Computers are built from connected switches that, like light switches, are either on or off. On is represented as 1, and off is
1 is called a ***bit***. 1011 is four bits. Eight bits, like 11000101, are called a ***byte***.

Humans represent information using characters and numbers like Z or 42. To present information that people can understa
need a way to represent characters and numbers using 0's and 1's.

**PARTICIPATION ACTIVITY** 2.1.1: Bits.

1) A 0 or 1 is called a _____ .

Check Show answer

2) Eight bits are called a _____ .

Check Show answer

3) 101100 has _____ bits.

Check Show answer

## Characters as bits: ASCII

A ***character*** is a letter (a, b, ..., z, A, B, ..., Z), symbol (!, @, #, ...), or single-digit number (0, 1, ..., 9). Basically, each item on a cc keyboard is a character (though more characters exist). Each character can be given a unique bit code.

***ASCII*** is a popular code for characters. ASCII stands for American Standard Code for Information Interchange, and was dev ASCII uses 7 bits per code, and has codes for 128 characters. Ex: Using ASCII, the letter Z would be stored in a computer as material inserts a space for readability, as in: 101 1010. Each bit code is sometimes written as an equivalent decimal numb Dec below), discussed later.

Table 2.1.1: ASCII bit codes for common characters.

| Bit code | Dec | Char | Bit code | Dec | Char | Bit code | Dec | Char |
|---|---|---|---|---|---|---|---|---|
| 010 0000 | 32 | space | 100 0000 | 64 | @ | 110 0000 | 96 | ` |
| 010 0001 | 33 | ! | 100 0001 | 65 | A | 110 0001 | 97 | a |
| 010 0010 | 34 | " | 100 0010 | 66 | B | 110 0010 | 98 | b |
| 010 0011 | 35 | # | 100 0011 | 67 | C | 110 0011 | 99 | c |
| 010 0100 | 36 | $ | 100 0100 | 68 | D | 110 0100 | 100 | d |
| 010 0101 | 37 | % | 100 0101 | 69 | E | 110 0101 | 101 | e |
| 010 0110 | 38 | & | 100 0110 | 70 | F | 110 0110 | 102 | f |
| 010 0111 | 39 | ' | 100 0111 | 71 | G | 110 0111 | 103 | g |
| 010 1000 | 40 | ( | 100 1000 | 72 | H | 110 1000 | 104 | h |
| 010 1001 | 41 | ) | 100 1001 | 73 | I | 110 1001 | 105 | i |
| 010 1010 | 42 | * | 100 1010 | 74 | J | 110 1010 | 106 | j |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 010 1011 | 43 | + | 100 1011 | 75 | K | 110 1011 | 107 | k |
| 010 1100 | 44 | , | 100 1100 | 76 | L | 110 1100 | 108 | l |
| 010 1101 | 45 | - | 100 1101 | 77 | M | 110 1101 | 109 | m |
| 010 1110 | 46 | . | 100 1110 | 78 | N | 110 1110 | 110 | n |
| 010 1111 | 47 | / | 100 1111 | 79 | O | 110 1111 | 111 | o |
| 011 0000 | 48 | 0 | 101 0000 | 80 | P | 111 0000 | 112 | p |
| 011 0001 | 49 | 1 | 101 0001 | 81 | Q | 111 0001 | 113 | q |
| 011 0010 | 50 | 2 | 101 0010 | 82 | R | 111 0010 | 114 | r |
| 011 0011 | 51 | 3 | 101 0011 | 83 | S | 111 0011 | 115 | s |
| 011 0100 | 52 | 4 | 101 0100 | 84 | T | 111 0100 | 116 | t |
| 011 0101 | 53 | 5 | 101 0101 | 85 | U | 111 0101 | 117 | u |
| 011 0110 | 54 | 6 | 101 0110 | 86 | V | 111 0110 | 118 | v |
| 011 0111 | 55 | 7 | 101 0111 | 87 | W | 111 0111 | 119 | w |
| 011 1000 | 56 | 8 | 101 1000 | 88 | X | 111 1000 | 120 | x |
| 011 1001 | 57 | 9 | 101 1001 | 89 | Y | 111 1001 | 121 | y |
| 011 1010 | 58 | : | 101 1010 | 90 | Z | 111 1010 | 122 | z |
| 011 1011 | 59 | ; | 101 1011 | 91 | [ | 111 1011 | 123 | { |
| 011 1100 | 60 | < | 101 1100 | 92 | \ | 111 1100 | 124 | \| |
| 011 1101 | 61 | = | 101 1101 | 93 | ] | 111 1101 | 125 | } |

| | | |
|---|---|---|
| 011 1110 | 62 | > |
| 011 1111 | 63 | ? |

| | | |
|---|---|---|
| 101 1110 | 94 | ^ |
| 101 1111 | 95 | _ |

| | | |
|---|---|---|
| 111 1110 | 126 | ~ |

**PARTICIPATION ACTIVITY** 2.1.2: ASCII bit codes (and decimal number equivalents).

| Type a character: | ASCII bit code: | ASCII number: |
|---|---|---|
| A | **1000001** | **65** |

**PARTICIPATION ACTIVITY** 2.1.3: ASCII.

1) What is the 7-bit code for a lower-case 'a'?

Check Show answer

2) What is the 7-bit code for a blank space?

Check Show answer

3) What two-letter word does this sequence of bits represents in ASCII? Pay attention to upper/lower case. Use

the above ASCII table.
1001000 1101001

Check Show answer

4) Suppose an email message has 500 characters. How many bits would a computer use to store that email, using ASCII code having bits per character?

bits

Check Show answer

**CHALLENGE ACTIVITY** 2.1.1: ASCII code.

Start

Convert the character to 7-bit ASCII code.
Ex: a is 1100001

k: Ex: 1100001

1 2 3 4

Check Next

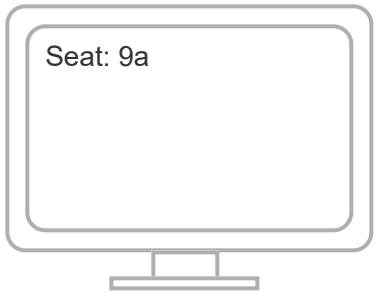
1
2
3
4

## Text is a sequence of character codes

Computers commonly deal with text, consisting of a sequence of characters. The computer stores each character's ASCII c successive locations in the computer's memory. Each location has at least enough bits (often more) to store an ASCII code

**PARTICIPATION ACTIVITY** 2.1.4: Characters are encoded and stored in the computer's memory.

Start 2x speed




Memory

| | |
|---|---|
| 101 0011 | S |
| 110 0101 | e |
| 110 0001 | a |
| 111 0100 | t |
| 011 1010 | : |
| 010 0000 | (space) |
| 011 1001 | 9 |
| 110 0001 | a |

**PARTICIPATION ACTIVITY** 2.1.5: Characters in a computer's memory.

Refer to the above animation.

1) How many characters are in the text displayed on the screen on the left side of the animation?

Check Show answer

2) How many memory locations are used to store that text?

Check Show answer

3) The first letter in the text is S. What bits are stored in the first memory location?

Check Show answer

4) What is the minimum number of bits that each memory location must be able to store in the example above?

Check Show answer

5) How many total bits are needed to store the text shown?

Check Show answer

## Encoding more characters: Unicode

***Unicode*** is another character encoding standard, published in 1991, whose codes can have more bits than ASCII and thus c over 100,000 items, such as symbols and non-English characters. Characters in Unicode are represented as a number, or ***c*** Unicode, the letter "H" is represented as U+0048. U+ means the character is encoded in Unicode, and 0048 is the correspor The code point is written in hexadecimal, which is discussed elsewhere.

Characters can range from U+0000 to U+10FFFF. The table below provides a very small subset of encodings.

Table 2.1.2: Unicode code points for control characters and basic Latin.

| Code point | Char | Code point | Char | Code point | Char | Code point | Char | Code point | Char | Code point | Cha |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0020 | space | 0030 | 0 | 0040 | @ | 0050 | P | 0060 | ` | 0070 | p |
| 0021 | ! | 0031 | 1 | 0041 | A | 0051 | Q | 0061 | a | 0071 | q |
| 0022 | " | 0032 | 2 | 0042 | B | 0052 | R | 0062 | b | 0072 | r |
| 0023 | # | 0033 | 3 | 0043 | C | 0053 | S | 0063 | c | 0073 | s |
| 0024 | $ | 0034 | 4 | 0044 | D | 0054 | T | 0064 | d | 0074 | t |
| 0025 | % | 0035 | 5 | 0045 | E | 0055 | U | 0065 | e | 0075 | u |
| 0026 | & | 0036 | 6 | 0046 | F | 0056 | V | 0066 | f | 0076 | v |
| 0027 | ' | 0037 | 7 | 0047 | G | 0057 | W | 0067 | g | 0077 | w |
| 0028 | ( | 0038 | 8 | 0048 | H | 0058 | X | 0068 | h | 0078 | x |
| 0029 | ) | 0039 | 9 | 0049 | I | 0059 | Y | 0069 | i | 0079 | y |
| 002A | * | 003A | : | 004A | J | 005A | Z | 006A | j | 007A | z |
| 002B | + | 003B | ; | 004B | K | 005B | [ | 006B | k | 007B | { |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 002C | , | 003C | < | 004C | L | 005C | \ | 006C | l | 007C | \| |
| 002D | - | 003D | = | 004D | M | 005D | ] | 006D | m | 007D | } |
| 002E | . | 003E | > | 004E | N | 005E | ^ | 006E | n | 007E | ~ |
| 002F | / | 003F | ? | 004F | O | 005F | _ | 006F | o | | |

UTF-8, UTF-16, and UTF-32 are encoding standards that indicate how the Unicode is stored. In the UTF-8 standards, charac using variable widths and range from one to four bytes. Whereas in the UTF-32 encoding, all characters are stored as a sing An application that converts the encoding to the final characters viewed by the end user must know which standard is utiliz pages, and other digital media frequently contain additional information, or metadata, to indicate how characters are stored may contain the tag <meta charset='utf-8'> to indicate that the UTF-8 Unicode standard is used to encode text.

**PARTICIPATION ACTIVITY** 2.1.6: Unicode.

1) An uppercase letter P is represented as _____ in Unicode.
   - U+0050
   - U+0070

2) U+0020 represents _____ .
   - an uppercase A
   - a space

3) Which text is represented by the following unicode?
   U+0061 U+0020 U+0062 U+0020 U+0063
   -

a b c

ABC

4) In Unicode, each character is stored as a 16-bit value.

True

False

**CHALLENGE ACTIVITY** 2.1.2: Unicode code point.

Start

Convert the character to 4-digit hexadecimal code point.
Ex: a = U+0061

R: U+ Ex: 0061

1 2 3 4

Check Next

This section provides a simple introduction to Unicode. We encourage the interested reader to Unicode Consortium for adc information on advanced topics and features.

Exploring further:

- Wikipedia: ASCII
- http://www.asciitable.com/
- Unicode 9.0 Character Code Charts

**Provide feedback on this section**