

Lecture Review

1. What is the signed + unsigned interpretation of 10000011?

Unsigned: $2^7 + 2^1 + 2^0 = 128 + 2 + 1 = 131$

Signed: $-2^7 + 2^1 + 2^0 = -128 + 2 + 1 = -125$

2. How many codepoints are encoded by the UTF-8 sequence?

1. 11010101 10111111 11101011 10110110 10000001

Two code points.

- 11010101 10111111: A 2-byte sequence (starts with 110..., followed by one 10... continuation byte) representing 1 codepoint (U+057F).

- 11101011 10110110 10000001: A 3-byte sequence (starts with 1110..., followed by two 10... continuation bytes) representing 1 codepoint (U+BD81).

3. What bit operations are needed to determine the integer codepoint of a multi-byte UTF-8 encoded character?

- Bitwise AND (&) with a bitmask: Clears the UTF-8 control prefixes (e.g., masking the lead byte with 0x1F, 0x0F, or 0x07 and continuation bytes with 0x3F) to isolate only the payload data bits.

- Bitwise Left Shift (<<): Shifts the extracted data bits left by multiples of 6 bits to align them into their correct positional significance within the resulting integer.

CSE - Bitwise OR (|): Combines the shifted bit fields from each byte together into a single unified integer codepoint.