Ch12 13 Fixed Floating Point Number Quiz ANS

1. In the Qm.n fixed-point notation (signed), what is the total number of bits used to represent the number?
A) m + n
B) $m + n + 1
C) m × n
D) m - n
ANS: B)
Slide 2 defines "Qm.n for signed fixed-point" as having m integer bits, n fractional bits, and 1 sign bit, totaling $m+n+1$ bits.

2. How is the value of a signed fixed-point number (Qm.n) calculated using the weighted bit method (where the MSB is the sign bit)?
A) The sign bit has a weight of $-2^{(m+1)}$.
B) The sign bit has a weight of -2^m.
C) The sign bit is ignored in calculation.
D) The sign bit has a positive weight of 2^m.
ANS: B)
Convert 10101.101 (Q4.3) by calculating $1 \times -2^4 + \cdots$. The MSB (sign bit) corresponds to -2^m (where $m = 4$ in the example).

3. In IEEE 754 Single Precision (FP32) format, what is the "Bias" value used for the exponent?
A) 127
B) 128
C) 255
D) 1023
ANS: A)
Bias = $2^7 - 1 = 127$ for single precision FP32.

4. Which of the following bit patterns represents the value "1.0" in IEEE 754 Single Precision format (Hexadecimal: 0x3F800000)?
A) Sign=0, Exponent=128, Fraction=0
B) Sign=1, Exponent=127, Fraction=0
C) Sign=0, Exponent=127, Fraction=0
D) Sign=0, Exponent=0, Fraction=1
ANS: C)
Decode 0x3F800000 (1.0). The exponent is 127. The formula is $(-1)^0 \times (1 + 0) \times 2^{\{127-127\}} = 1 \times 1 \times 2\text{^}0 = 1.0$.

5. When converting a floating-point number f to a fixed-point number with n fractional bits, what is the first step in the calculation?
A) Divide f by 2^n.
B) Add the bias to f.
C) Multiply f by 2^n.
D) Invert the bits of f.
ANS: C)
Slide 8 ("Convert Float to Fixed-point UQ4.12") shows the first step as "Calculate $f \times 2^{12}$".

6. In IEEE 754, what does an exponent field of all 1s (e.g., 255 for FP32) and a non-zero fraction field represent?
A) Positive Infinity ($+\infty$)
B) Negative Infinity ($-\infty$)
C) Zero
D) NaN (Not a Number)
ANS: D)
Slide 22 ("Special Values") shows that Exponent=255 (all 1s) with a Fraction $\neq 0$ is "NaN". If the fraction were 0, it would be Infinity.

7. What is the primary difference between "Normalized" and "Subnormalized" numbers in IEEE 754?
A) Subnormalized numbers have an implicit leading bit of 0, while Normalized numbers have an implicit 1.
B) Subnormalized numbers use a bias of 255.
C) Normalized numbers cannot represent negative values.
D) Subnormalized numbers are used only for values greater than 1.
ANS: A)
Slide 23 contrasts Normalized $(-1)^S \times (1 + F)$ ... with Subnormalized $(-1)^S \times (0 + F)$ ....

8. According to the document, how does Fixed-Point representation differ from Floating-Point regarding precision?
A) Fixed-Point precision decreases as the number's magnitude increases.
B) Floating-Point numbers are evenly distributed across the entire range.
C) Fixed-Point numbers are evenly distributed, meaning precision is fixed.
D) Fixed-Point generally has a much larger range than Floating-Point.
ANS: C)
Slide 27 ("Tradeoff between Range and Precision") notes that for Fixed-Point, "Numbers are evenly distributed... Precision is fixed," whereas for Floating-Point, "Precision decreases as the magnitude increases."

9. What are the "bf16" and "bf8" formats primarily used for?
A) High-precision scientific simulations requiring 64-bit accuracy.
B) AI and machine learning to trade slightly lower accuracy for speed and efficiency.
C) Legacy banking systems using COBOL.
D) Storing high-fidelity audio files.
ANS: B)
bf16 and bf8 are used in AI and machine learning for efficient computation... where accuracy can be slightly sacrificed.

10. If you decode the binary sequence 1 10000011 11111... (Sign=1, Exponent=131), what is the resulting sign and exponent value before bias adjustment?
A) Sign: Positive, Exponent: 4
B) Sign: Negative, Exponent: 131 (Actual exponent 131 - 127 = 4)
C) Sign: Negative, Exponent: 3
D) Sign: Positive, Exponent: -4
ANS: B)
Slide 15 decodes 0xC1FF0000. Sign bit 1 means Negative. Exponent 10000011 is 131. The actual exponent is 131 - 127 = 4.