计算机系统概论（2024 秋）作业 1 参考答案

1. 在所有由五个 “1” 和三个 “0” 组成的 8 位二进制整数（补码形式）中，最小的数是_10001111_，最大的数是_01111100_（答案用二进制表示）。（每个 5 分，共 10 分）

2. 已知 [X] 补= 0x2024，[Y] 补= 0xAE77，则 [X+Y] 补=_0xCE9B_，[X-Y] 补=_0x71AD_（X、Y 的数据位宽均为 16 位，计算结果用 16 进制的补码表示）。（每个 5 分，共 10 分）

3. BF16 是一种 16 位浮点数类型（符合 IEEE 浮点数标准），常见于大模型混合精度训练。BF16 的 exp 位数是 8，frac 位数是 7，符号位数是 1，其所能表示的最大的非规格化数的 exp 是00000000，frac 是1111111。2024（十进制数）的 exp 是10001001，frac 是 1111101（请用 0、1 位串表示答案）。（每个 5 分，共 20 分）

4. 假设存在一种 9 位浮点数（符合 IEEE 浮点数标准），符号位数是 1，exp 位数是 4，frac 位数是 4。其数值被表示为 $V = (-1)^S \times M \times 2^E$ 形式。请在下表中填空。（每个 3 分，全对 20 分）

描述	Binary	M	E	Value
5.0	010010100	$1.010_2 = 1.25_{10}$	2	5.0
最小的大于 0 的浮点数	000000001	$0.0001_2 = 0.0625_{10}$	-6	2^{-10}

5. FP16 是另一种 16 位浮点数（符合 IEEE 浮点数标准），也常见于大模型混合精度训练。FP16 的 exp 位数是 5，frac 位数是 10，符号位数是 1。某同学对该格式的一个数 x 执行了（整数的）按位右移操作，得到了 80.5(0 10101 0100001000)。若右移操作按有符号数执行（算术右移），原来的数可能是_不存在_，若右移操作按无符号数执行（逻辑右移），原来的数可能是_-$\frac{97}{2^{11}}$, -$\frac{1553}{2^{15}}$ or -$\frac{97}{2048}$, $-\frac{1553}{32768}$ _（列出所有情况或填入 “不存在”，数可以用小数或分数来表示，必须精确）。（每个答案 5 分，共 10 分）

6. 给定相同的字长（例如 16 位），（每个问号 5 分，共 20 分）

   (a) 能表示的定点数个数多还是浮点数个数多？为什么？

– 定点数中，不同的 01 串所表示所对应的数完全不同。

– 浮点数中：

(a) 规格化数不同的 01 二进制表示所对应的数不同。

(b) 非规格化数不同的 01 二进制表示所对应的数不同，正负除外（两个不同的 01 串表示同一个数）。

(c) 存在两个 01 串分别表示正负无穷。

(d) 存在多个不同的 01 串表示 NaN。

综上，相同字长能表示的定点数较多。

(b) BF16 格式表示的浮点数个数多还是 FP16 格式表示的浮点数个数多？在模型训练中如果遇到梯度爆炸（模型参数数值大小趋向 ∞）或者梯度消失（模型参数数值大小趋向 0），希望通过改变浮点数格式来缓解数值问题时，应该选用 BF16 还是 FP16？由第一问，主要比较 NaN 所占 01 串数量；NaN 数量取决于 frac 长度，因为 BF 的 frac 更短，所以 BF 数量更多；BF16 表示范围更大，遇到提到消失和爆炸应该用 BF16。

7. 使用不超过 4 次位运算或加减运算完成整数运算 $y = x \times 85$（允许引入临时变量，不需要考虑溢出的情况）。（每个步骤 5 分，共 10 分） 85=1010101, t=(x«2+x), y=(t«4+t).