**Question 2 (20 points):** At the Hot Chips Conference in Cupertino, CA, in August 2017, Microsoft announced that its Project Brainwave would leverage Field-Programmable Gate Arrays (FPGAs) to deliver real-time Artificial Intelligence through Deep Neural Networks. Microsoft disclosed that this new hardware would use a Microsoft custom 8-bit floating point format (ms-fp8). While no details have been disclosed for this format, earlier publications have proposed FP8 with one bit for sign, four bits for exponent and three bits for mantissa. Assuming that Microsoft made this design decision and that no representation is reserved for special values such as NaN and $\pm\infty$, we can assume the following FP8 format definition.

| 7 | 6   3 | 2   0 |
|---|---|---|
| S | *biasedexponent* | *fraction* |

$$N = \begin{cases} (-1)^S \times 0.\mathit{fraction} \times 2^{-6} & \text{if } \mathit{biasedexponent} = 0 \\ (-1)^S \times 1.\mathit{fraction} \times 2^{\mathit{biasedexponent}-7} & \text{if } 0 < \mathit{biasedexponent} \leq 15 \end{cases}$$

1. (**6 points**) What is the largest positive number that can be represented in this FP8 format? Provide both the binary representation and the decimal value.

   Binary representation: `0111 1111`
   Decimal value: $1.111 \times 2^{15-7} = 1.111 \times 2^8 = 111100000 = 2^9 - 2^5 = 2^5 \times (2^4 - 1) = 32 \times 15 = 480$

2. (**6 points**) What is the smallest positive non-zero number that can be represented in this FP8 format? Provide both the binary representation and the decimal value.

   Binary representation: `0000 0001`
   Decimal value: $0.001 \times 2^{-6} = 1.0 \times 2^{-9} = \frac{1}{2^9} = \frac{1}{512} = 0.001953125 = 1.953125 \times 10^{-3}$

3. (**8 points**) Let $A =$ `0x41` and $B =$ `0x39` be the binary representation of two numbers in this FP8 format. What is the representation of $A + B$, expressed in hexadecimal? Assume that the hardware that performs this addition has a guard, a round, and a sticky bit. Also assume that the round-to-the-closest-even strategy is used when needed.

   $A =$ `01000001` $= 1.001 \times 2^{8-7} = 1.001 \times 2^1$
   $B =$ `00111001` $= 1.001 \times 2^{7-7} = 1.001 \times 2^0 = 0.1001 \times 2^1$

```
  1.001 0 0 0

+ 0.100 1 0 0

------------------

  1.101 1 0 0
```

   Must round up to round to the nearest even. Thus $A + B = 1.110 \times 2^1$
   Thus, $\mathit{biasedexponent} - 7 = 1 \Rightarrow \mathit{biasedexponent} = 8$
   $A + B =$ `0100 0110` $=$ `0x46`