ECE 487 Homework I

1) Fixed-point signed integer

Range of numbers $\Rightarrow$ $-2^{n-1} \leq X \leq 2^{n-1} - 1$

a) n_bit Signed integer fixed_point

if $n=3$ $\quad -2^{+2} \leq X \leq 2^{2} - 1 \quad A=4 \;\; B=3$

$-4 \leq X \leq 3 \quad A+B+1=8$ numbers can be represented.

It means 2^n number can be represented by n_bit Signed integer fixed_point.

b) 2n_bit signed integer fixed_point.

2^{2n} numbers can be represented.

c) Ratio of b to a is $\frac{2^{2n}}{2^{n}} = 2^{n}$

2) Fixed-point n_bit signed fractional

Range of numbers $\Rightarrow$ $-1 \leq X \leq (1 - 2^{-(n-1)})$

a) for n; Resolution is $\frac{1}{2^{n-1}}$, Precision is $\frac{1}{2^{n-1}} \times 100\%$

for 2n; Resolution is $\frac{1}{2^{2n-1}}$ $\quad \frac{1}{2^{2n-1}} \Big/ \frac{1}{2^{n-1}} = \frac{1}{2^{n}}$ times precision is increased.

b) for n+k; Resolution is $\frac{1}{2^{n+k-1}}$ $\quad \frac{1}{2^{n+k-1}} \Big/ \frac{1}{2^{n-1}} = \frac{1}{2^{k}}$ times presion is increased.

c) for $n=4$. If we want to improve precision by a factor of at least 6, I choose to improve by 8, because 8 is a factor of 2 something. At $n=4$ resolution is $2^{-(n-1)} = \frac{1}{8}$ if we want 8 times better resolution it should be $\frac{1}{64}$

$2^{-(?-1)} = \frac{1}{64} \qquad ? = 7$

As a result, if we want 8 times better resolution, we should add at least 3 bit to n. improving resolution 8 times is also same improving precision.

3) This question is same with question1.

a) n-bit signed integer fixedpoint format can represent. 2^n numbers.

b) n+1 bit represent 2^{n+1} numbers.

c) $2^{n+1}/2^n \Rightarrow 2$ with every bit, numbers doubled.

d) if we want quadruple range of numbers.

$2^{n+x}/2^n \Rightarrow 4$ $x=2$ we should add 2 bits.

4) In n-bit floating-point format representation, we use only mantissa bits for resolution, precision.

a) $\frac{1}{2^k}$ is resolution, $\frac{1}{2^k} \times \%100$ is precision.

k=number mantissa bits.

b) if $k=4$ $\frac{1}{16} \times 100 \Rightarrow \% 6,25$

if $k=5$ $\frac{1}{32} \times 100 \Rightarrow \% 3,125$

if $k=6$ $\frac{1}{64} \times 100 \Rightarrow \% 1.5625$

4 times better precision.

With every bit precision betters two times (2 added bit factor)