III. Floating Point Representation

- Floating Point numbers contain the following components
  1) Mantissa
  2) Mantissa sign (optional): Some method to allow a signed mantissa such as: a) an explicit sign bit, or b) another bit to allow a signed format with the same positive range
  3) Exponent
  4) Exponent sign (optional): Some method to allow a signed mantissa such as: a) an explicit sign bit, or b) another bit to allow a signed format with the same positive range
  5) Exponent base. The value is normally fixed and not explicitly provided. Common values are 2, a power-of-2, or 10 (used for financial applications).

Floating Point Basics

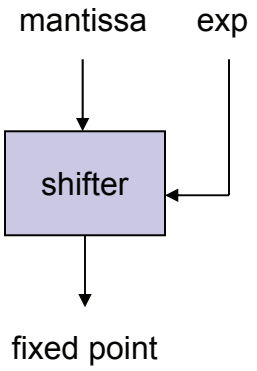
- Normally write: *mantissa* * $2^{exponent}$
- Alternately, *sign* * *mantissa* * $2^{exponent}$
- In hardware, we normally operate on, transmit, and save only the mantissa and exponent:
  (`MMMMMMM, EEEE`)
- Normalized floating point numbers contain no extra (useful) bits at the MSB of the mantissa
  - Example:

`00010 * 2^0`	not normalized, or "denormalized"
`01000 * 2^(-2)`	normalized, with 2's comp. mantissa
`10000 * 2^(-3)`	normalized, with unsigned mantissa

Floating Point → Fixed Point Conversion

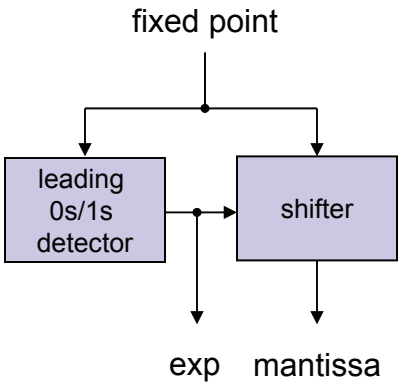
- If the *exp* is unsigned, the shifter shifts only to the left
- If the *exp* is signed, the shifter must shift to the left and right
- Example:

```
    01011. * 2^2
    01011. << 2
 0101100.
```



Fixed Point → Floating Point Conversion

- Leading 0s/1s detector finds the optimum place to begin selecting bits for the mantissa
- Common pitfall: If the *mantissa* is signed, its sign bits must be maintained!



Floating Point

- Fixed-to-float conversion example (*positive* input)
  - Input: 8-bit 2's complement (signed) integer
    Output: 4-bit 2's complement (signed) mantissa

    a) integer mantissa

```
0 0 0 0 1 1 0 0.  →   0 1 1 0. * 2^(001)   % 2^1
        12        =      6      *  2^1
```

    b) fractional "0.4 format" mantissa

```
0 0 0 0 1 1 0 0.  →  .0 1 1 0  * 2^(101)   % 2^(5)
        12        =    0.375   *   2^5
```


Floating Point

- Fixed-to-float conversion example (*negative* input)
  - Input: 8-bit 2's complement (signed) integer
    Output: 4-bit 2's complement (signed) mantissa

    a) integer mantissa

```
1 1 0 1 0 0 0 1.  →   1 0 1 0. * 2^(011)   % 2^3
      -47         =     -6     *  2^3
```

    b) fractional "2.2 format" mantissa

```
1 1 0 1 0 0 0 1.  →   1 0.1 0  * 2^(101)   % 2^(5)
      -47         =     -1.5   *  2^5
```