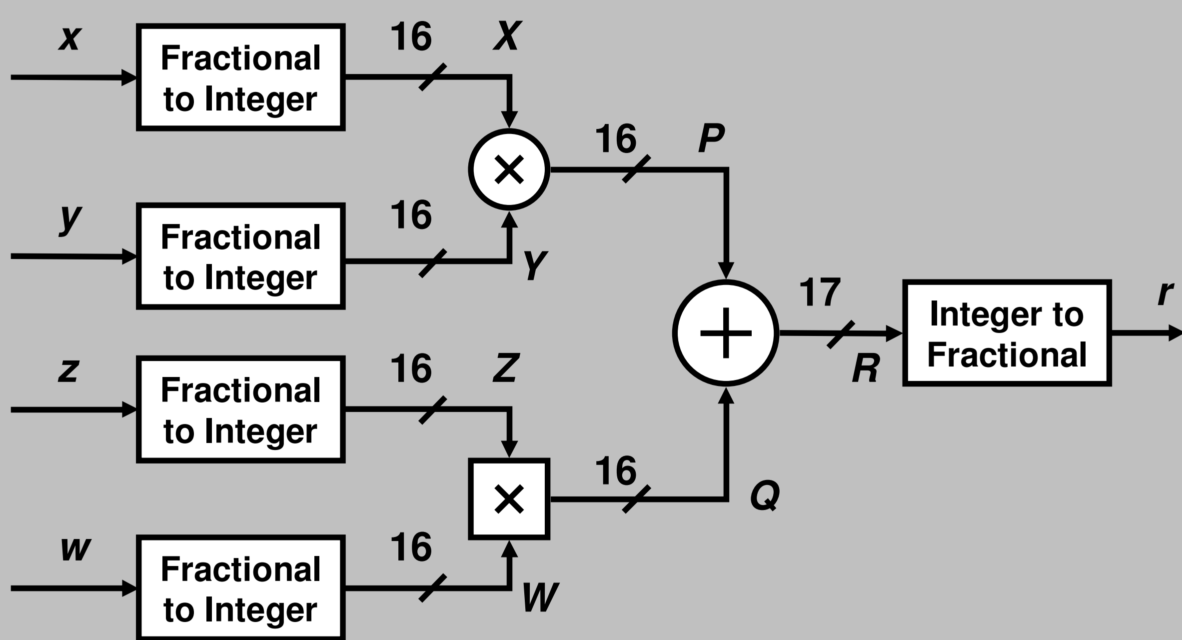
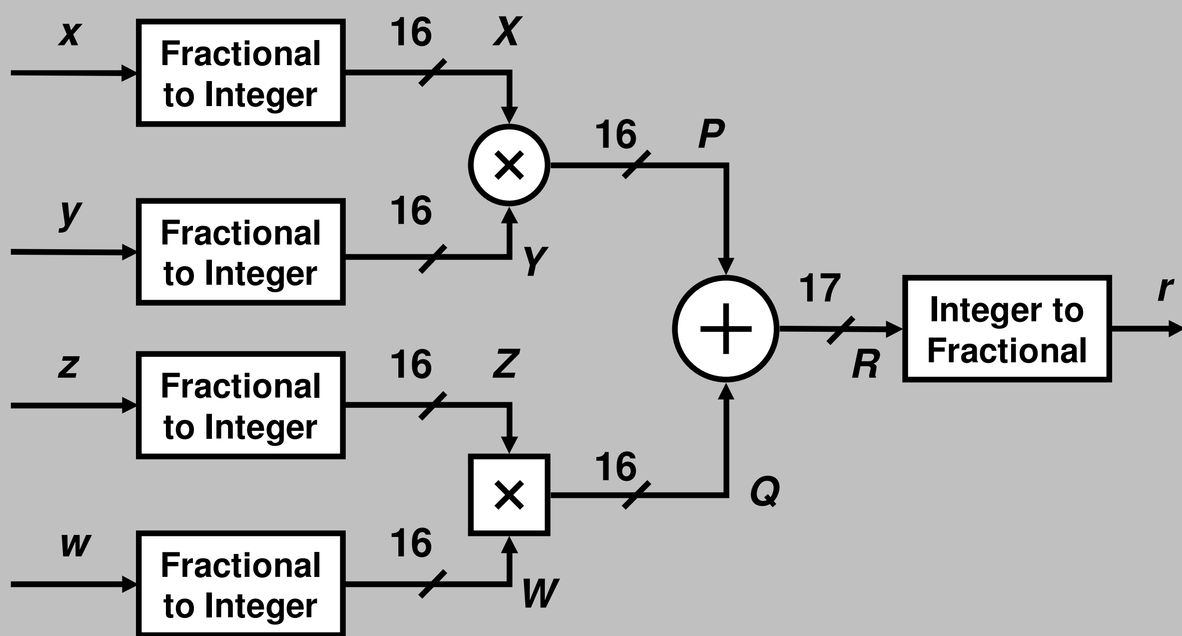
Fixed-point arithmetic

- **$-0.5 < x < +0.5$**
- X is a 16-bit fixed-point representation of x
- **0.5** is represented as $\mathbf{2^{15}}$
- x is represented as $X = x \cdot 2^{15} / 0.5 = x \cdot 2^{16}$



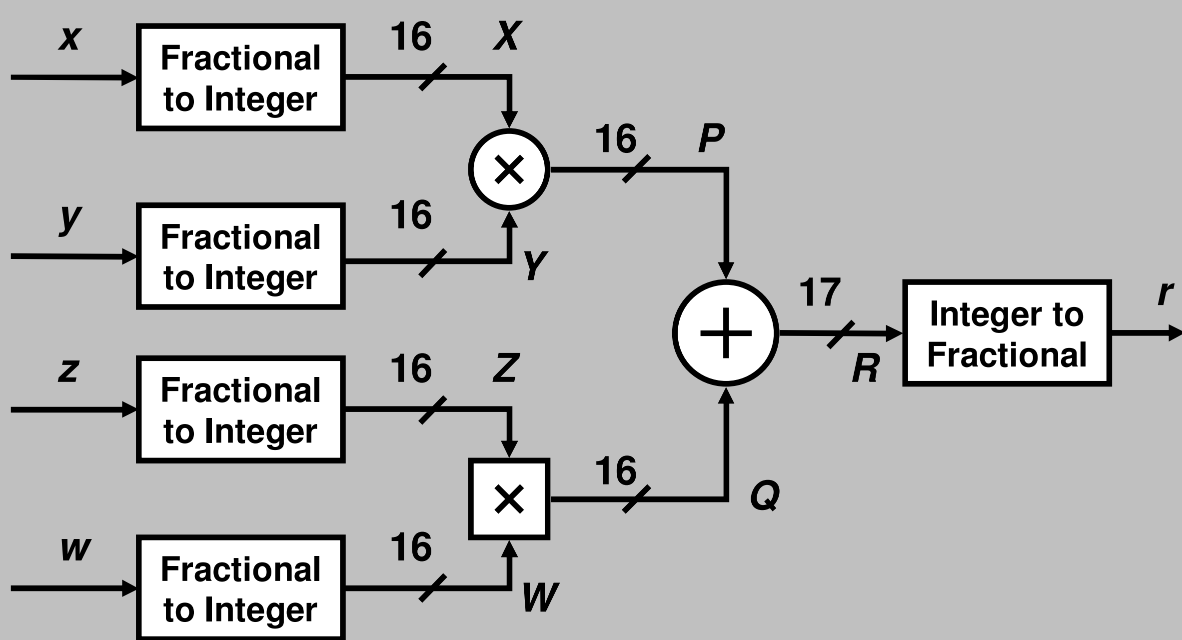
Fixed-point arithmetic

- **-1.0 < y < +1.0**
- Y is a 16-bit fixed-point representation of y
- **1.0** is represented as 2^{15}
- y is represented as $Y = y \cdot 2^{15} / 1.0 = y \cdot 2^{15}$



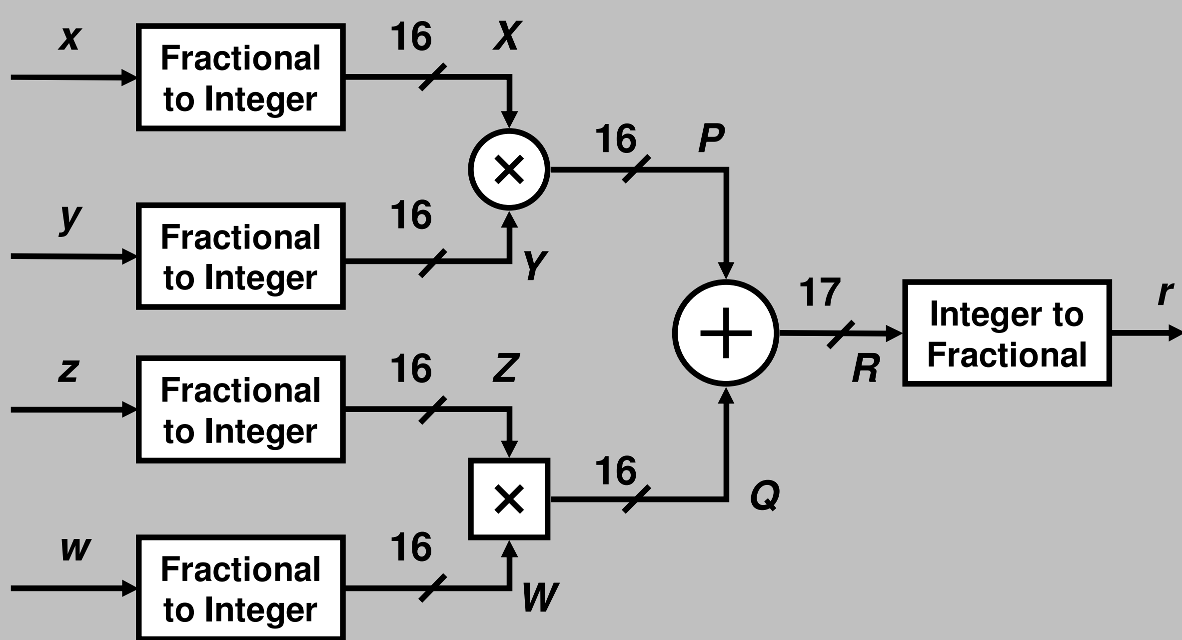
Fixed-point arithmetic

- **-2.0 < *z* < +2.0**
- ***Z*** is a 16-bit fixed-point representation of ***z***
- **2.0** is represented as $\mathbf{2^{15}}$
- ***z*** is represented as $Z = z \cdot 2^{15} / 2.0 = z \cdot 2^{14}$



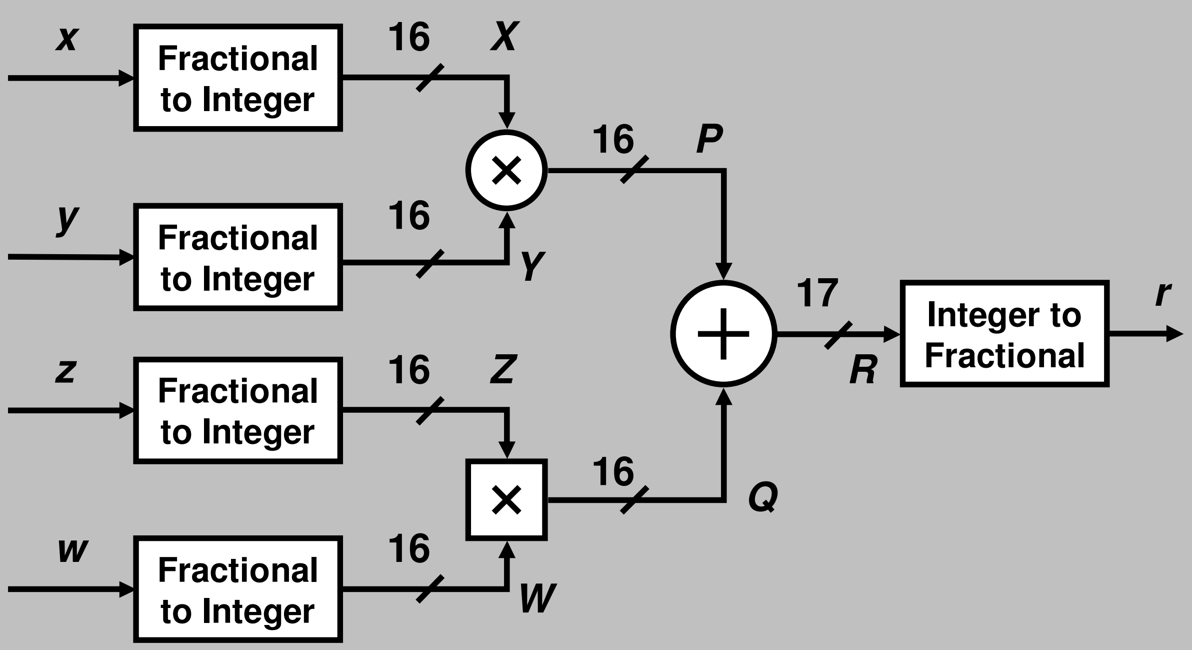
Fixed-point arithmetic

- **$-4.0 < w < +4.0$**
- ***W*** is a 16-bit fixed-point representation of ***w***
- **4.0** is represented as $\mathbf{2^{15}}$
- ***w*** is represented as $W = w \cdot 2^{15} / 4.0 = w \cdot 2^{13}$



Fixed-point arithmetic

- $x = -0.3 \rightarrow X = -0.3 \cdot 2^{16} = -19661 = \text{B333 h}$
- $y = 0.7 \rightarrow Y = 0.7 \cdot 2^{15} = 22938 = \text{599A h}$
- $z = 1.6 \rightarrow Z = 1.6 \cdot 2^{14} = 26214 = \text{6666 h}$
- $w = -3.1 \rightarrow W = -3.1 \cdot 2^{13} = -25395 = \text{9CCD h}$



- ***X***, ***Y***, ***Z***, ***W***: they are all 16-bit signed integers

Fixed-point arithmetic

- Fractional multiplier: the extra bit is the product's LSbit

  P_{dp}: Product **P** in **d**ouble-**p**recision
- $\boldsymbol{P_{dp} = X \cdot Y \cdot 2 = \text{B333 h} \cdot \text{599A h} \cdot 2 = \text{CA3D0F5C h}}$
- P_{dp} is a 32-bit signed integer
- The scale factor of P_{dp} is $\mathbf{2^{16} \cdot 2^{15} \cdot 2 = 2^{32}}$
- $\boldsymbol{p_{dp} = x \cdot y \cdot 2 = [(-0.3 \cdot 2^{16}) / 2^{16}] \cdot [(0.7 \cdot 2^{15}) / 2^{15}] \cdot 2 / 2}$

  $\boldsymbol{p_{dp} = [X / 2^{16}] \cdot [Y / 2^{15}] \cdot 2 = [X \cdot Y \cdot 2] / 2^{32} = P_{dp} / 2^{32}}$
- P is a 16-bit signed integer: $\boldsymbol{P = P_{dp} >> 16 = \text{CA3D h}}$

  von Neumann and truncation
- $\boldsymbol{p = (P_{dp} / 2^{32}) \cdot (2^{16} / 2^{16})}$

  $\boldsymbol{p = (P_{dp} >> 16) / 2^{16} = \text{CA3D h} / 2^{16} = -13763 / 2^{16}}$

  von Neumann and truncation

Fixed-point arithmetic

- Integer multiplier: the extra bit is the product's MSbit

  $\boldsymbol{Q_{dp}}$: Product **Q** in **d**ouble-**p**recision
- $\boldsymbol{Q_{dp} = Z \cdot W = \text{6666 h} \cdot \text{9CCD h} = \text{D85227AE h}}$
- $\boldsymbol{Q_{dp}}$ is a 32-bit signed integer
- The scale factor of $\boldsymbol{Q_{dp}}$ is $\boldsymbol{2^{14} \cdot 2^{13} = 2^{27}}$
- $\boldsymbol{q_{dp} = z \cdot w = [(1.6 \cdot 2^{14}) / 2^{14}] \cdot [(-3.1 \cdot 2^{13}) / 2^{13}]}$

  $\boldsymbol{q_{dp} = [Z / 2^{14}] \cdot [W / 2^{13}] = [Z \cdot W] / 2^{27} = Q_{dp} / 2^{27}}$
- $\boldsymbol{Q}$ is a 16-bit signed integer: $\boldsymbol{Q = Q_{dp} >> 16 = \text{D853 h}}$

  von Neumann and truncation
- $\boldsymbol{q = (Q_{dp} / 2^{27}) \cdot (2^{16} / 2^{16})}$

  $\boldsymbol{q = (Q_{dp} >> 16) / 2^{11} = \text{D853 h} / 2^{11} = -10157 / 2^{11}}$

  von Neumann and truncation

Fixed-point arithmetic

- Addition: $r = p + q = P / 2^{16} + Q / 2^{11}$
- It is not allowed to increase the word-length, thus P will be right-shifted over 4 positions:

  $r = (P / 2^{16}) \cdot (2^5 / 2^5) + Q / 2^{11} = [(P >> 5) + Q] / 2^{11}$

  von Neumann

  $r = (\mathrm{FE52\ h} + \mathrm{D853\ h}) / 2^{11}$

- The addition generates an extra bit of **carry**, thus both operands will be represented over 17-bit fields by extending their sign bit:

  $r = (\mathrm{1FE52\ h} + \mathrm{1D853\ h}) / 2^{11} = \mathrm{1D6A5\ h} / 2^{11}$

  $R = \mathrm{1D6A5\ h} = -10587$, which is a 17-bit signed integer

- $r = -10587 / 2^{11} = -5.16943$

Fixed-point arithmetic

<u>Common mistake</u>:

- **Q = D853 h / 2^{11}**
- **2^{11}** specifies the **position of the binary point**
- **2^{11}** has nothing to do with the number of bits for representing the fixed-point value **Q**

D				**8**				**5**				**3**			
1	1	0	1	1	0	0	0	0	1	0	1	0	0	1	1

Binary point