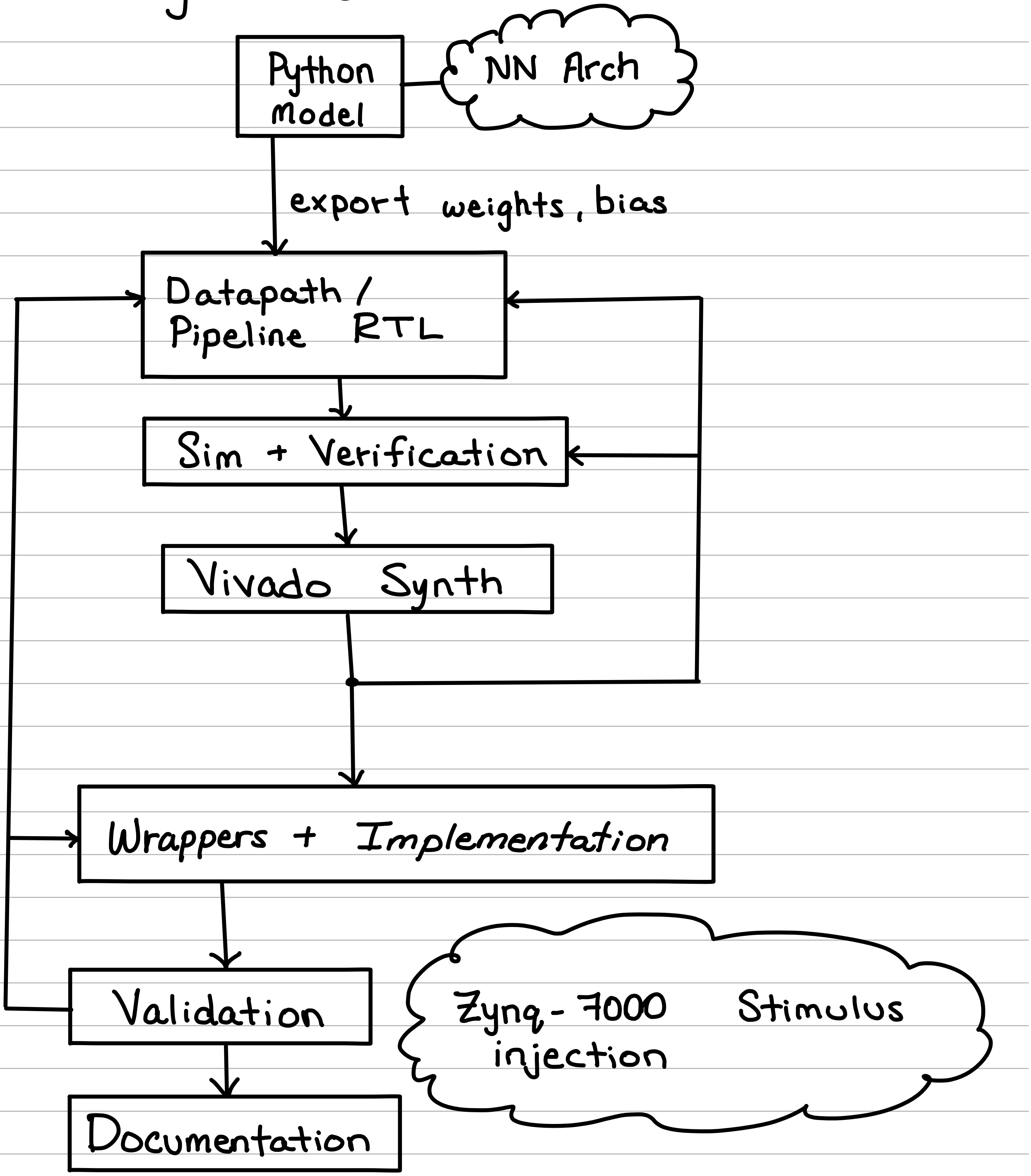
# MNIST - Accelerator :

→ Design Flow



# MNIST Neural Network Design

Input : 28 x 28 , 8-bit grayscale images flattened to a 784 vector.

Output : Size 10 vector representing strength for each digit.

Architecture : 784 → 500 → 10

Linear (784, 500)
ReLU (500)
Linear (500, 10)

Linear Layers : $y = xA^T + b$

Layer #1 : $A = (500, 784)$
$b = (500)$

Layer #2 : $A = (500, 10)$
$b = (10)$

Neurons : $y = xw + b$

Layer #1 : $x = (784)$
$w = (784)$
$b = (1)$
$y = (1)$

Layer #2 : $x = (500)$
$w = (500)$
$b = (1)$
$y = (1)$

# RTL Design :

## Key Problems :

→ Data Representation (Bits? Fixed?)
→ Data Operations (Add, Sub, Mult)
→ Functions
- ReLU (max(0, z))
  + Neuron / Layer Design
- Linear ($y = xA^T + b$)
  + Neuron / Layer Design

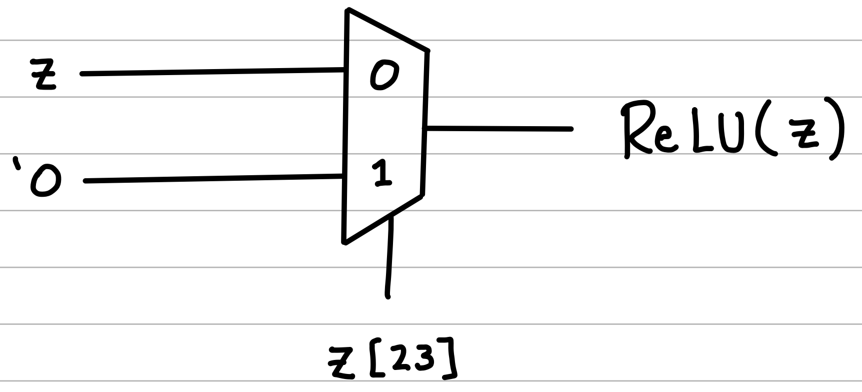
→ Pipelining / Datapath

## 24-Bit Fixed Point :

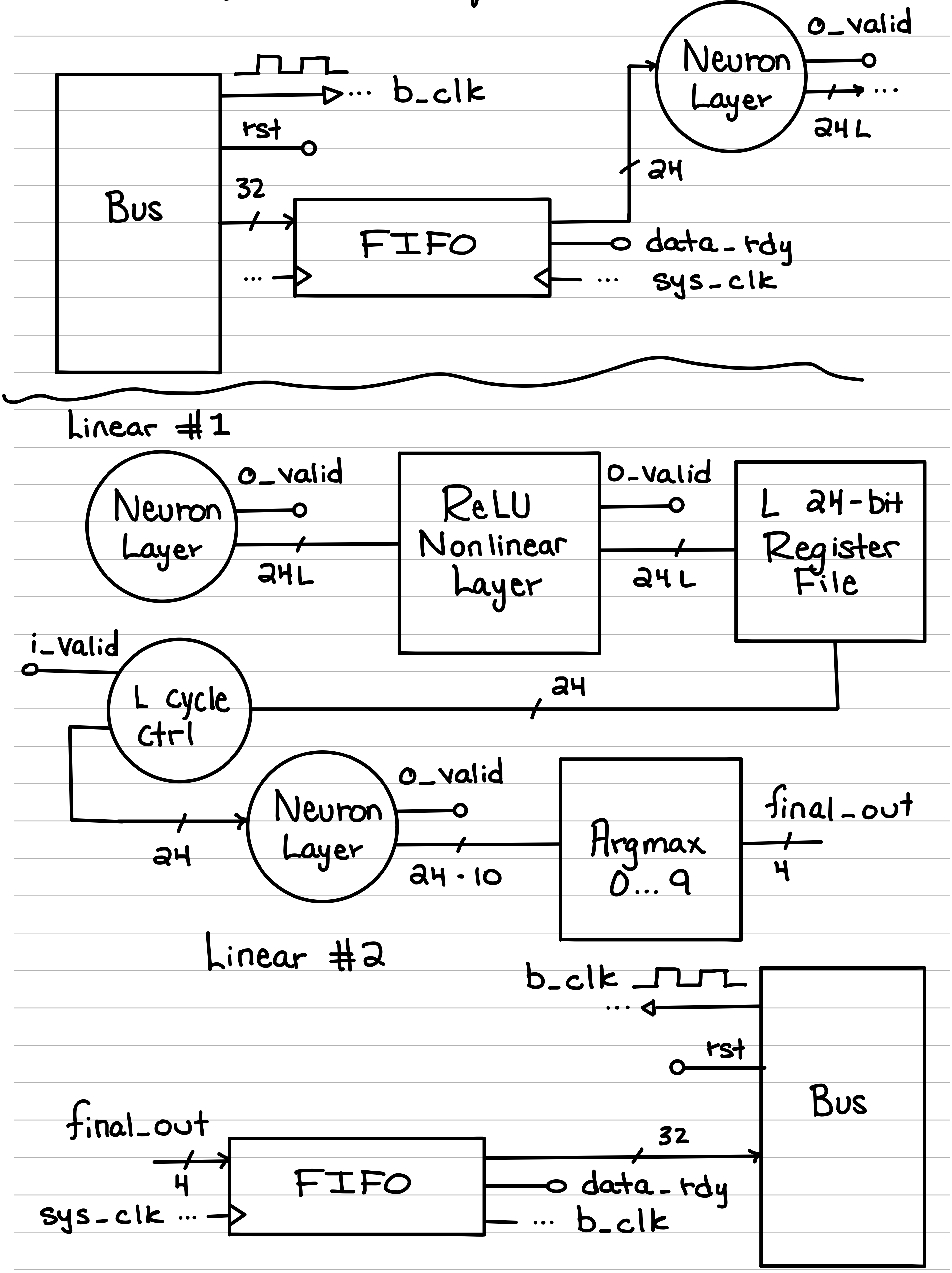
| 9-bit | 15-bit |
|---|---|
| Integer | Decimal |

- Add, Sub are normal
- Mult must >> by # decimal (30)

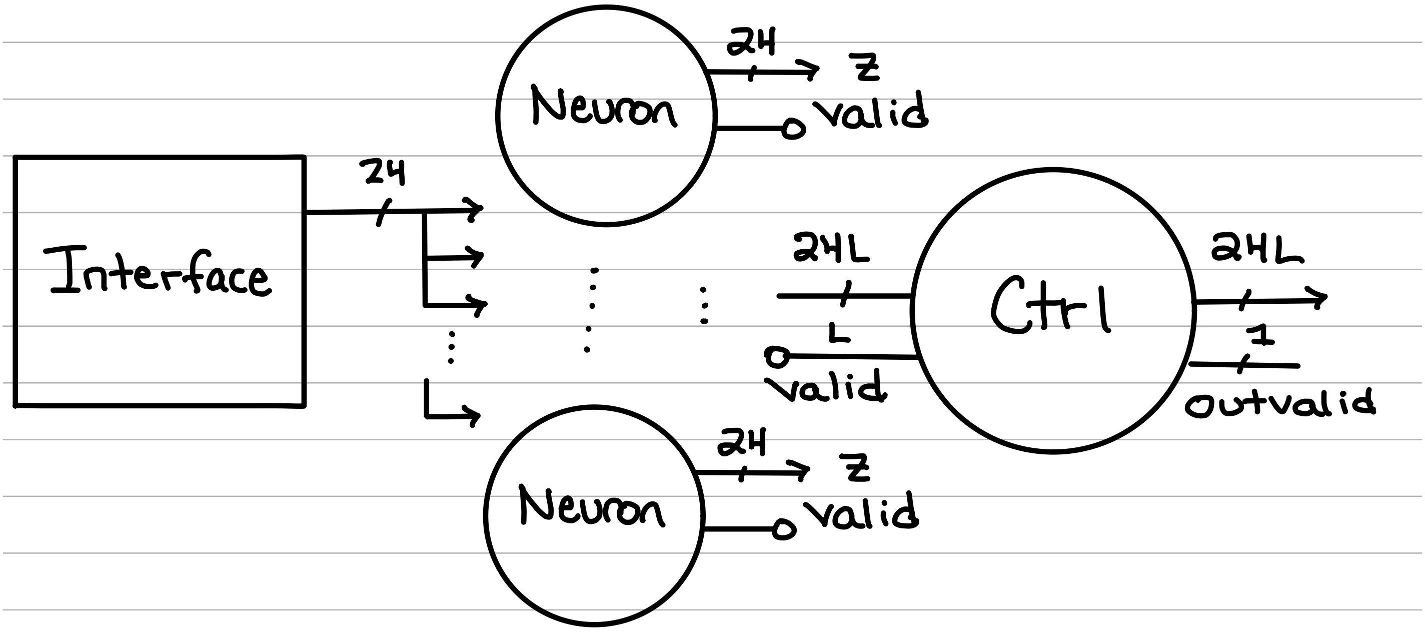
## ReLU Cell RTL :



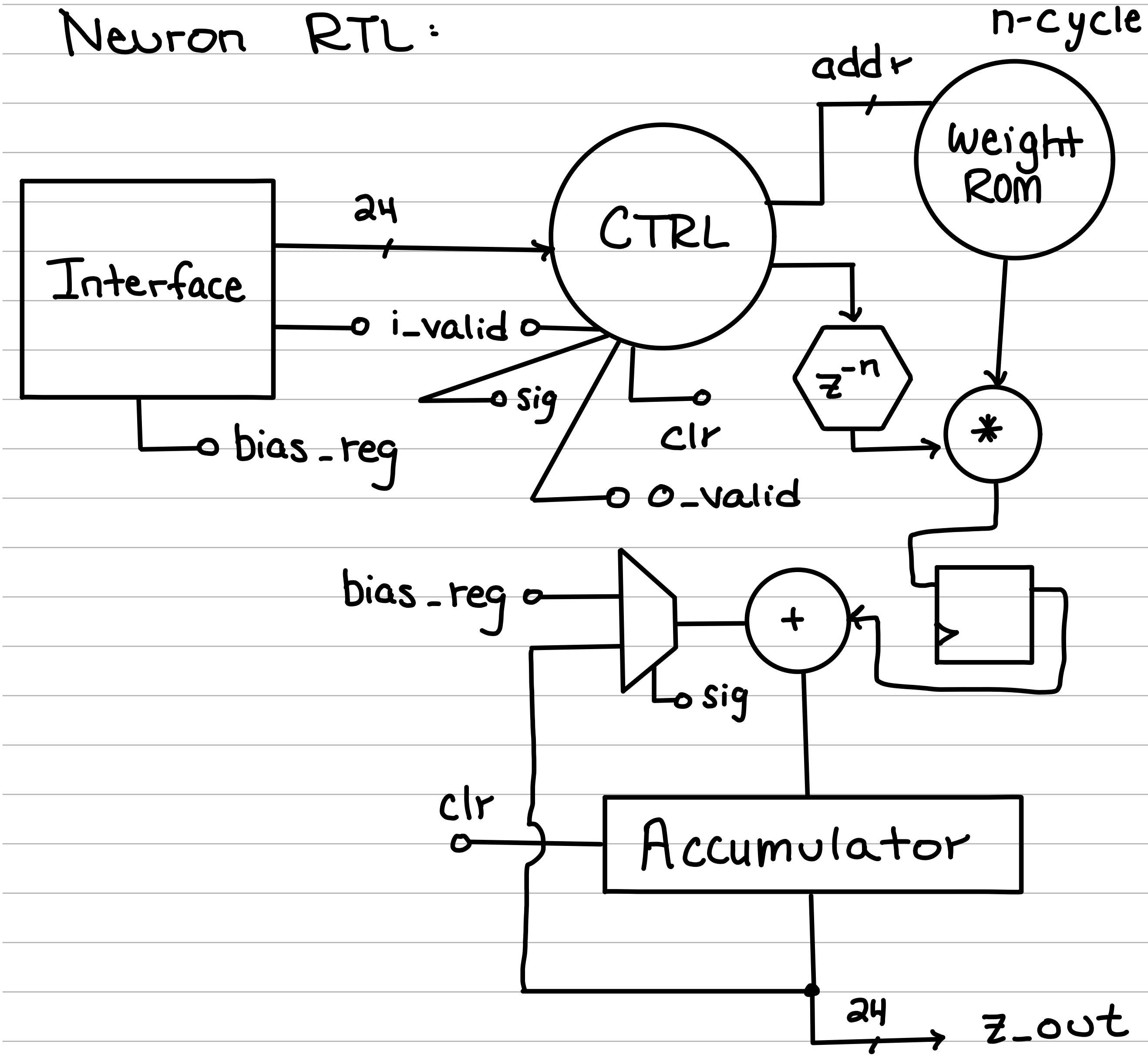
Model RTL : L = # Layers



## Linear Layer RTL:



## Neuron RTL:


n-cycle
addr
Weight ROM
Interface
24
CTRL
i_valid
sig
bias_reg
clr
o_valid
z^{-n}
*
bias_reg
+
sig
clr
Accumulator
24
z_out