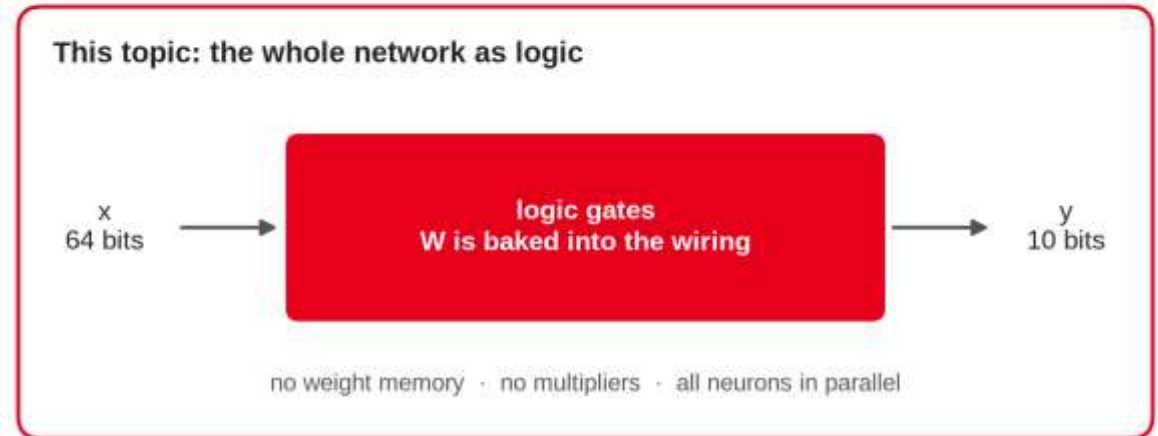
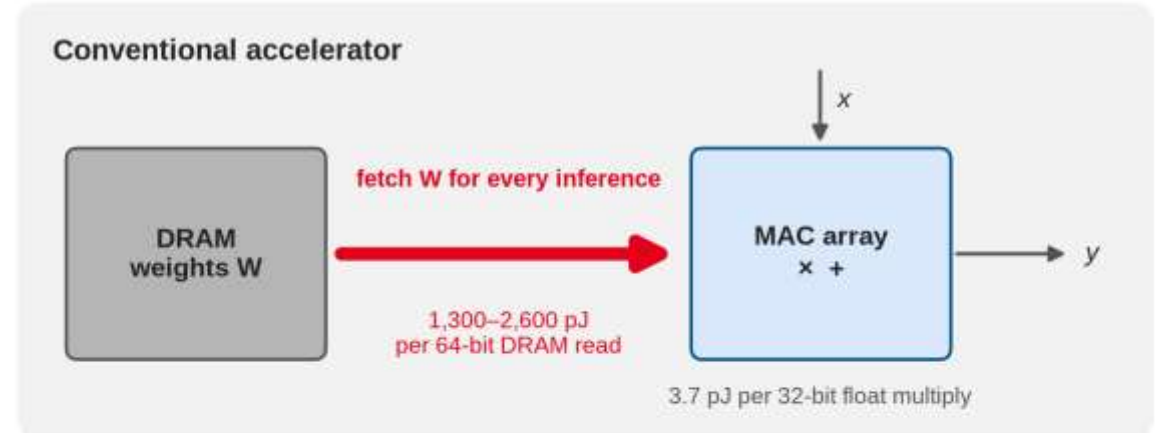


FROM NEURAL NETWORK TO GATE-LEVEL NETLIST

M.SC BAOYIN HUA

MOTIVATION

- Inference cost is dominated by data movement, not arithmetic
 - One 64-bit DRAM read $\approx 350\text{--}700\times$ the energy of a 32-bit float multiply (45 nm)
- **Idea: if every input and output of a neuron can be represented as 0 or 1, the neuron is a Boolean function**
 - No arithmetic computation
 - No stored weights
 - No external computation resources
- Reference result (NullaNet, ASP-DAC 2019), two hidden layers of an MNIST MLP:
 - **6,400 $\times$** fewer memory accesses
 - **97 $\times$** less logic than parallel 32-bit MACs

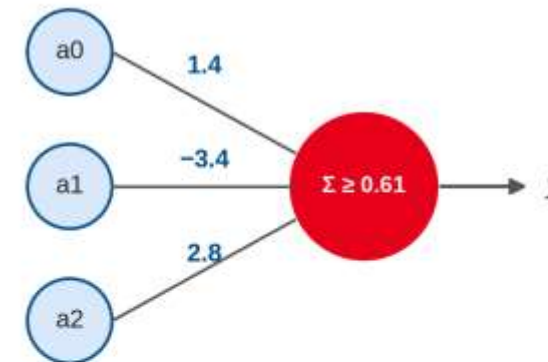


FROM NEURON TO LOGIC GATES

- **Binary inputs → finite but huge input space**
 - With n binary inputs, a neuron can only ever see 2^n patterns
- **Finite input space → truth table**
 - Record only what the network actually sees
- **Truth table → logic gates**
 - Sum of products (an OR of ANDs); AND, OR and NOT can build any Boolean function
 - Using **Incompletely Specified Function** to minimize boolean function
- **A layer is a multi-output Boolean function; a network is layers in series → one combinational circuit**

➤ The weights disappear into the wiring

(a) Neuron

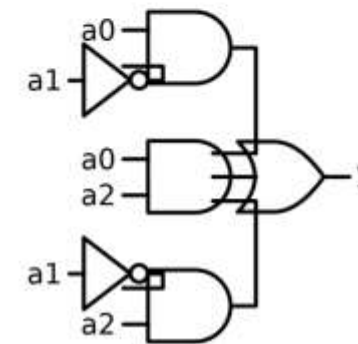


(b) Truth table

a_0	a_1	a_2	$\Sigma w \cdot a$	y
0	0	0	0.0	0
0	0	1	2.8	1
0	1	0	-3.4	0
0	1	1	-0.6	0
1	0	0	1.4	1
1	0	1	4.2	1
1	1	0	-2.0	0
1	1	1	0.8	1

blue rows: $y = 1$ (ON-set)

(c) Minimize → logic gates



$$y = a_0 \cdot \neg a_1 + a_0 \cdot a_2 + \neg a_1 \cdot a_2$$

YOUR TASK

▪ Challenges

- Binary input and output
- Minimal Boolean expression from Truth Table
- Remaining one-hot output

▪ Tasks and Target

- Compare accuracy: model vs. netlist
- Analyse netlist cost: area, critical-path delay, power

▪ Think about

- Can the netlist be more accurate than the original model? Why?
- Compare using one truth table for each layer and using a truth table for the entire neural network?
- Versus MAC-based hardware: what do we save, what do we pay?
- Neuron has different inputs but identical outputs. Is it possible and acceptable? Why?
- On a new input the output bits may not be one-hot. Why and What then?

