DIGITAL IMPLEMENTATION OF ARTIFICIAL NEURAL NETWORKS

CADENCE DESIGN CONTEST 2018 (SLOT 3:45 TO 4:15)

SHOYEB KHAN

B.E(Electronics And Communication)

Dr. RAGHURAM SRINIVAS

Associate Professor

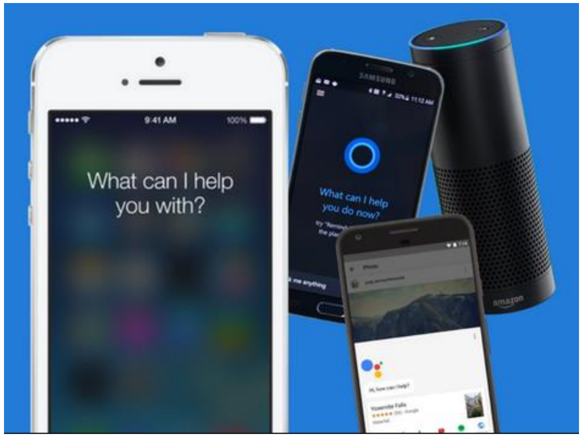
Dept. of E & C,

RIT, Bangalore

RAMAIAH
Institute of Technology

The First Challenge : Model Size

1.Models are getting larger, but the devices are getting smaller



Voice Assistant

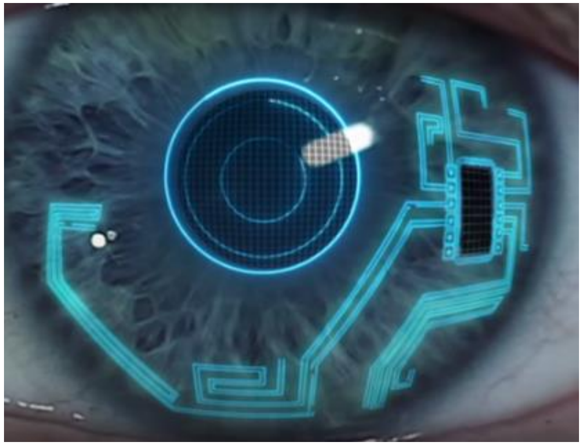


Image Recognition

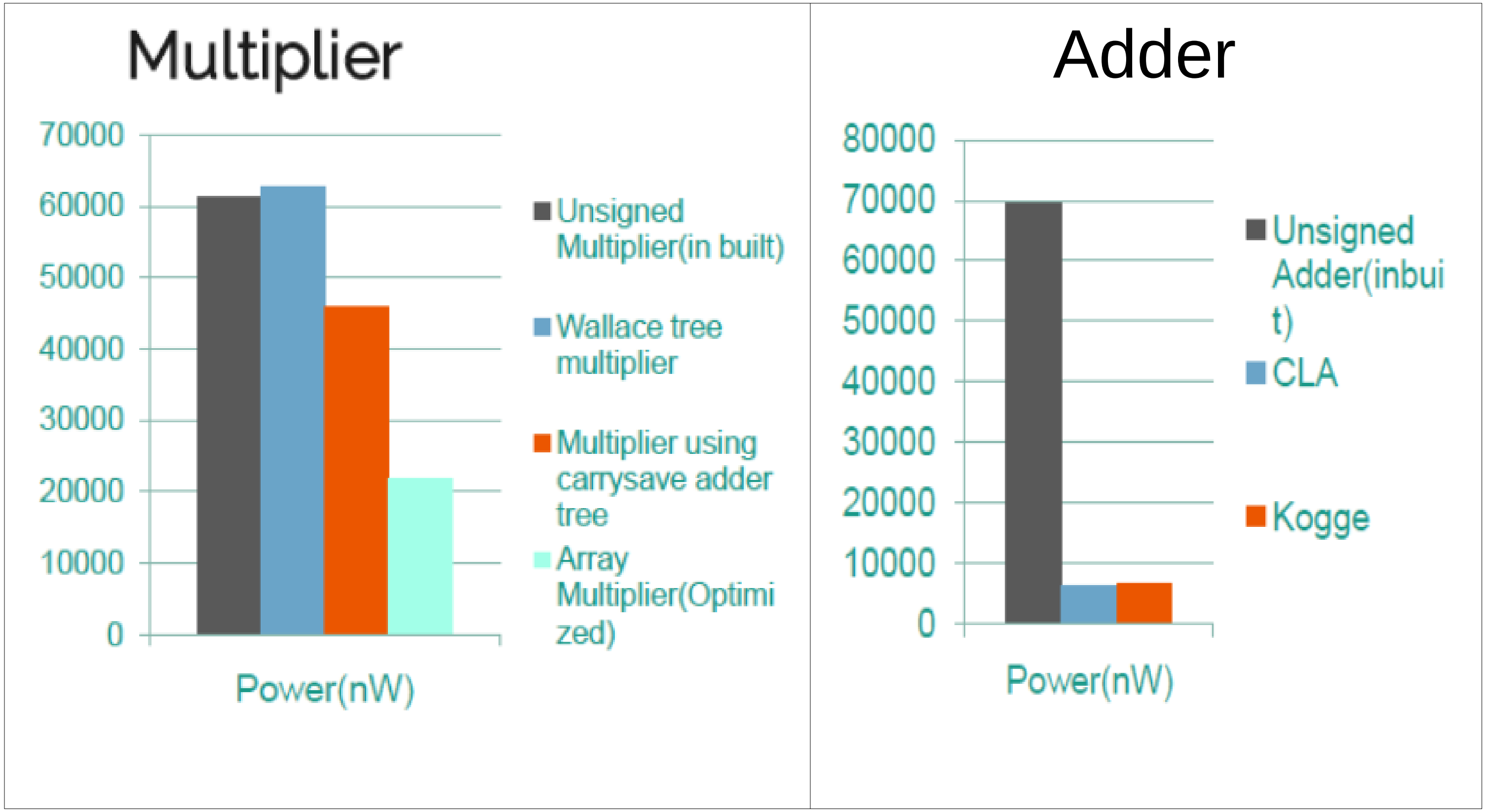
Analysis

Issues in hardware implementation:

a. Limited number of fan-in.
b. Operators are area greedy.
c. Weight storage cannot be included on the FPGA.
d. Activation functions also consumes a significant part.

The Second Challenge : Energy Efficiency

Single neuron power consumption : 137432 nW (Cadence® Encounter® RTL)
Largest NN used for image recognition has 650000 neurons.[A Krizhevsky et al.]
Avg. Power consumption for ImageNet : 89.3 Watts
IPhone 6 uses 10.5 Watts-hour for charging.

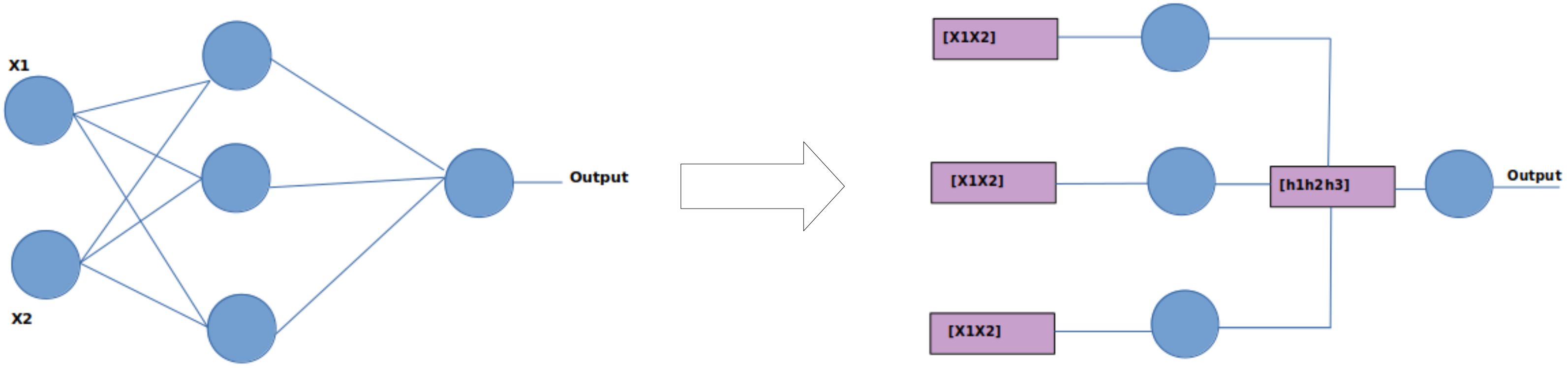


I consume a lot of Power!!!

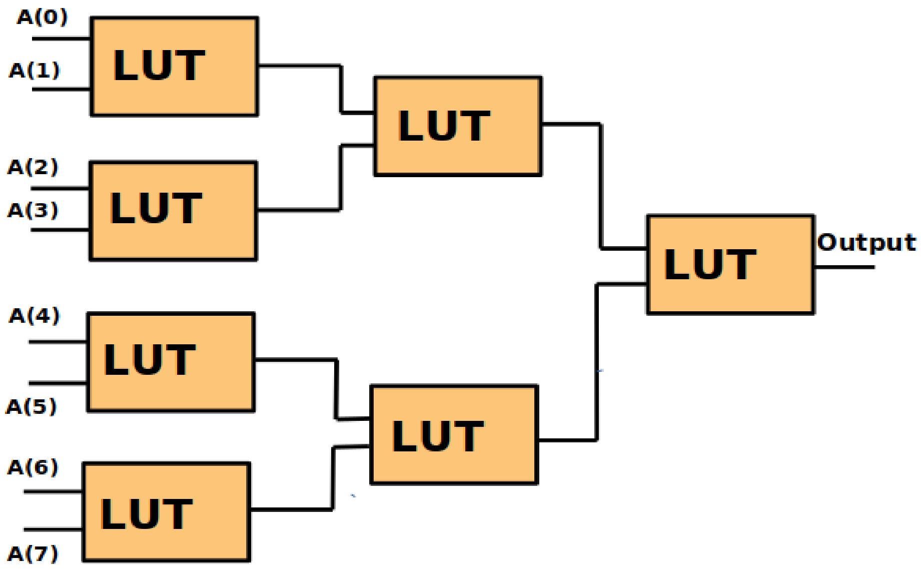
Single Artificial Neuron

PROPOSED SOLUTIONS

a. Reducing the number of interconnections.



b. Redesign of the Activation Function.



c. Shift the computation inside the memory

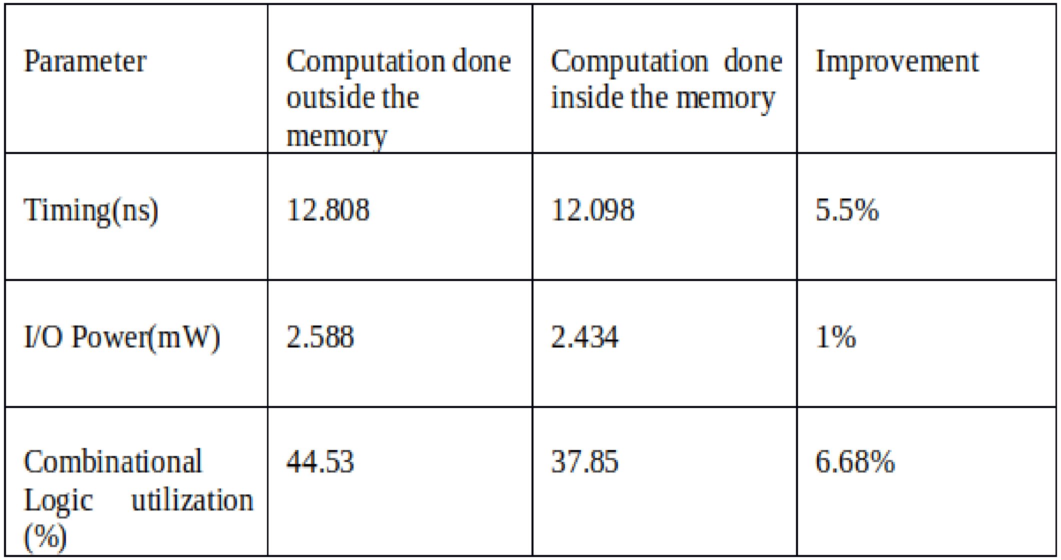
Where is the energy consumed ??
1. The operators are area as well as power greedy.
2. Weights need memory references => more energy consumed .

Operation	Energy(pJ)
16 bit ADD	0.1
16 bit MULT	3.1
16 bit Register file	1
16 SRAM cache	5
16 bit DRAM Memory	640

1 DRAM = 1000 (X and +)

[Han et al.]

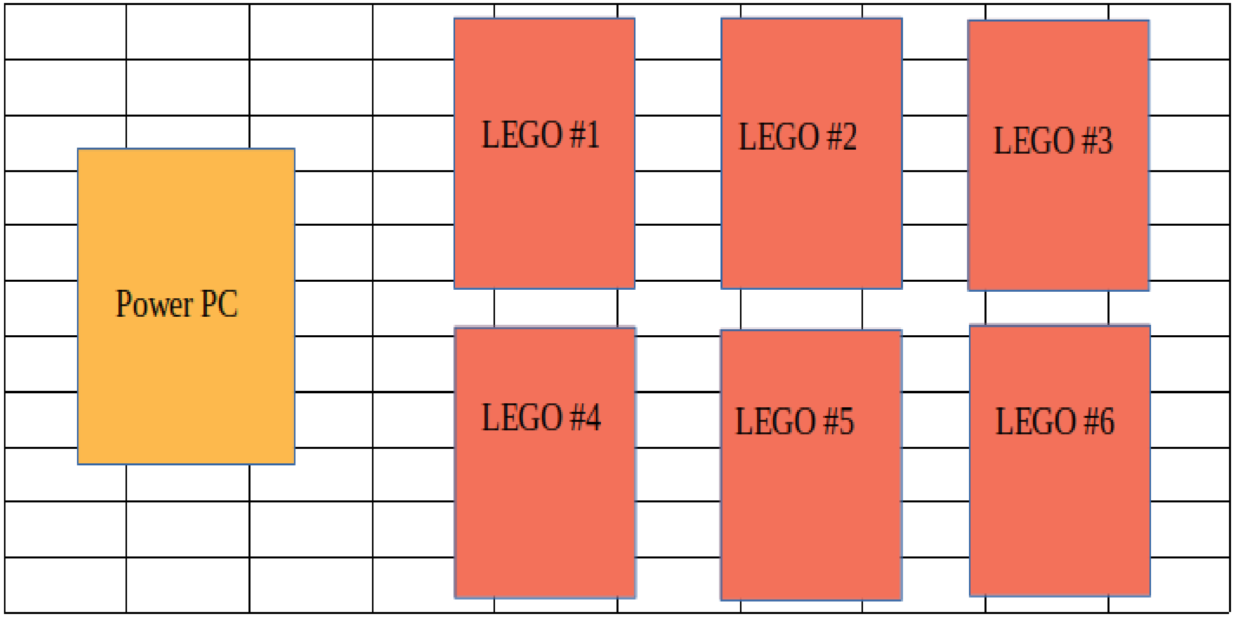
Improvement Result due to computation in the memory.



Parameter	Computation done outside the memory	Computation done inside the memory	Improvement
Timing(ns)	12.808	12.098	5.5%
I/O Power(mW)	2.588	2.434	1%
Combinational Logic utilization (%)	44.53	37.85	6.68%

c. Asynchronous Design

a. Demonstrated ability for async. circuits to consume power only on demand.
b. LEGO^{TM} approach.
c. Object-oriented approach to hardware.
d. Avoid clock distribution problems.
e. Natural way to describe systems with lots of concurrency.



Power PC

LEGO #1

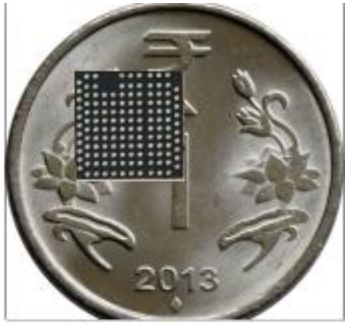
LEGO #2

LEGO #3

LEGO #4

LEGO #5

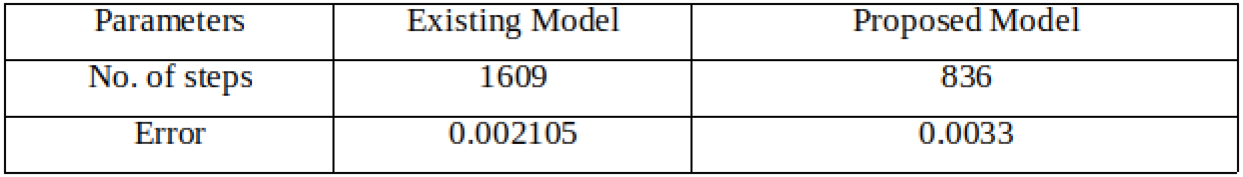
LEGO #6



To get to the next level in performance/Watt innovation at the AI chip level should include:
1.low precision computing
2.resistive computing

RESULTS

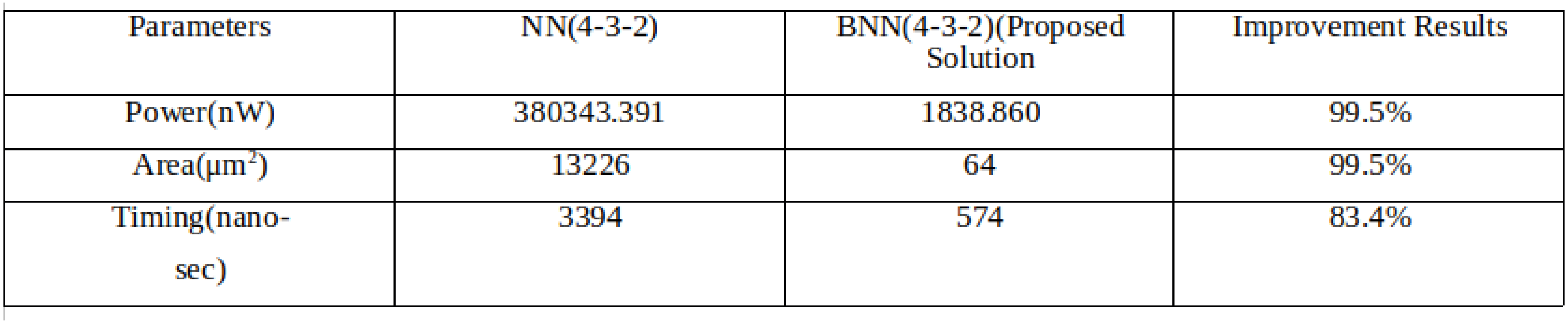
A.) Software version of the model



Parameters	Existing Model	Proposed Model
No. of steps	1609	836
Error	0.002105	0.0033

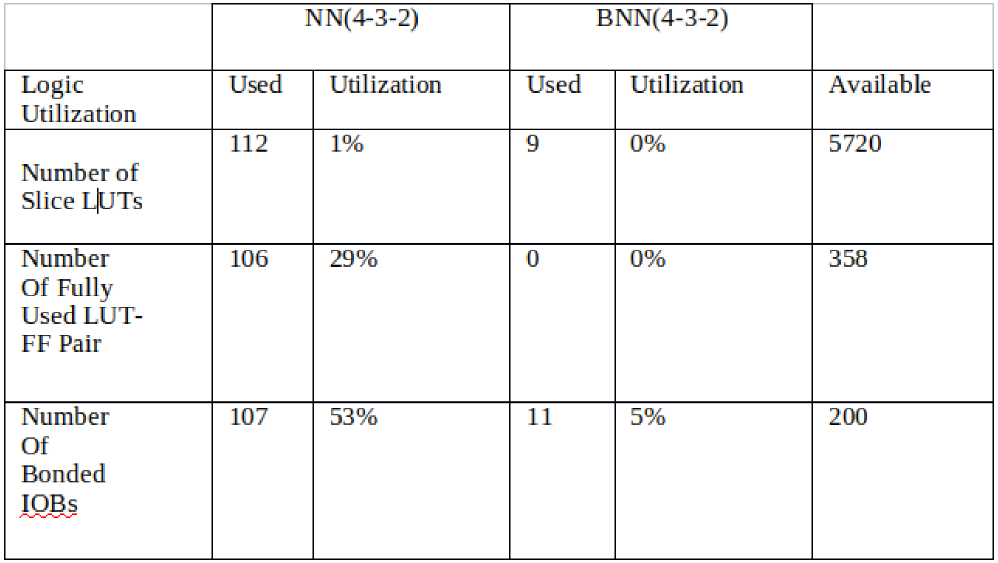
B.) Hardware Version of the model

a.) ASIC Implementation

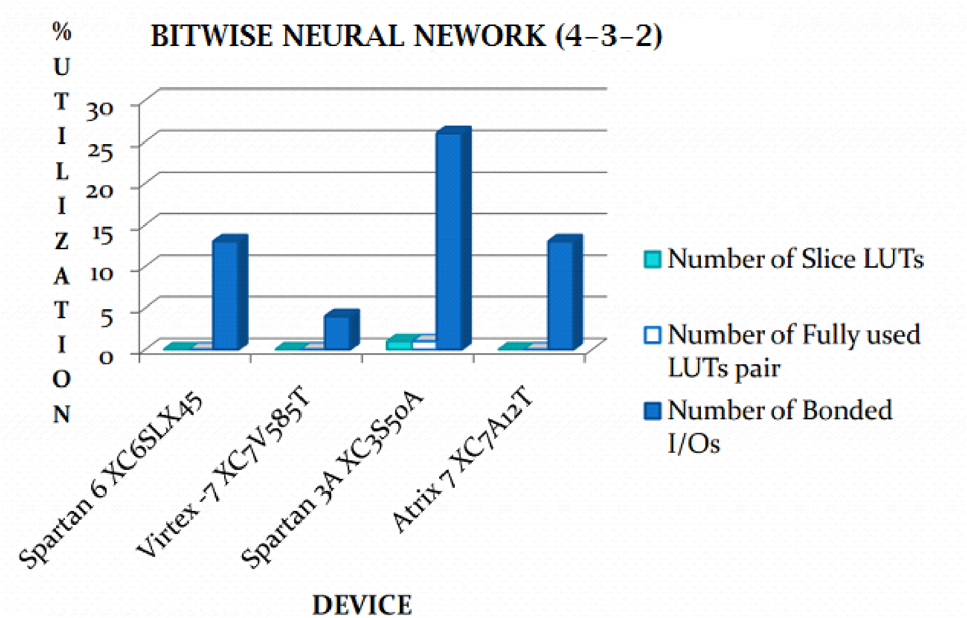


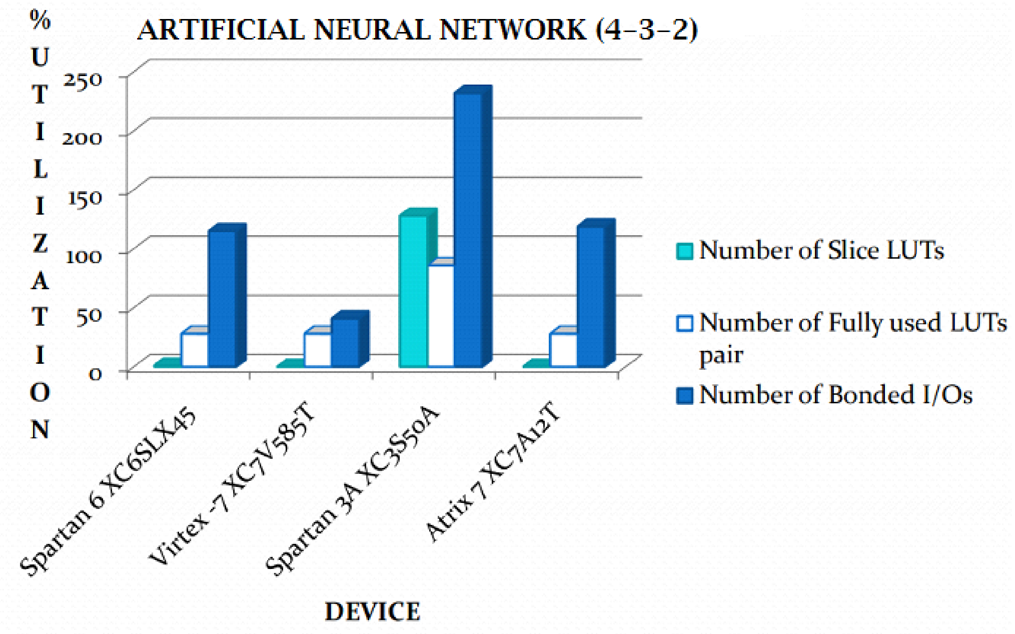
Parameters	NN(4-3-2)	BNN(4-3-2)(Proposed Solution	Improvement Results
Power(nW)	380343.391	1838.860	99.5%
Area(µm^2)	13226	64	99.5%
Timing(nano-sec)	3394	574	83.4%

b.) FPGA Implementation

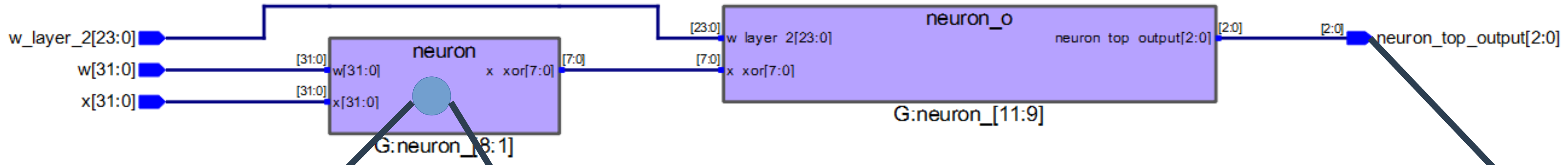


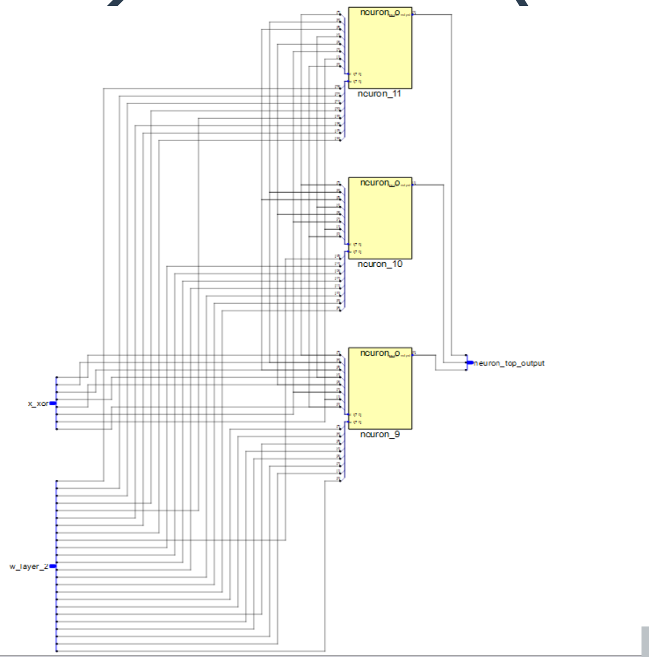
	NN(4-3-2)		BNN(4-3-2)		
Logic Utilization	Used	Utilization	Used	Utilization	Available
Number of Slice LUTs	112	1%	9	0%	5720
Number Of Fully Used LUT-FF Pair	106	29%	0	0%	358
Number Of Bonded IOBs	107	53%	11	5%	200

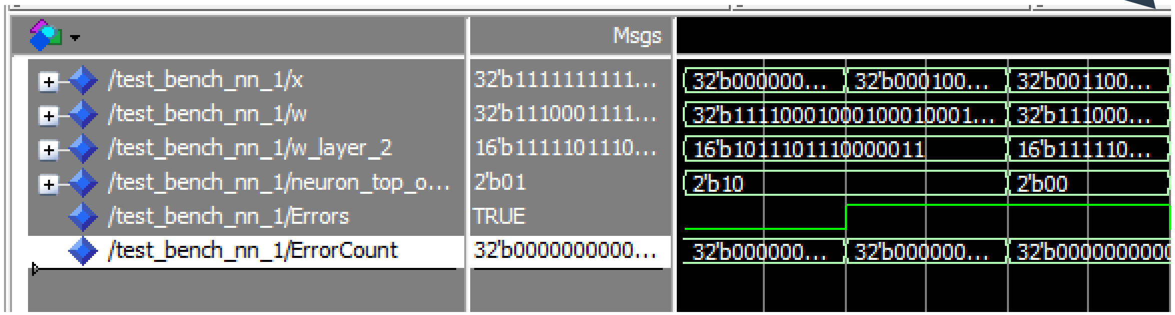




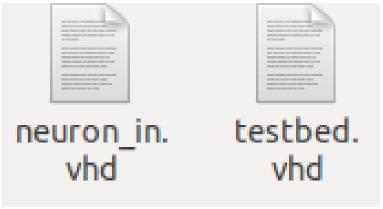
DESIGN AND SIMULATION





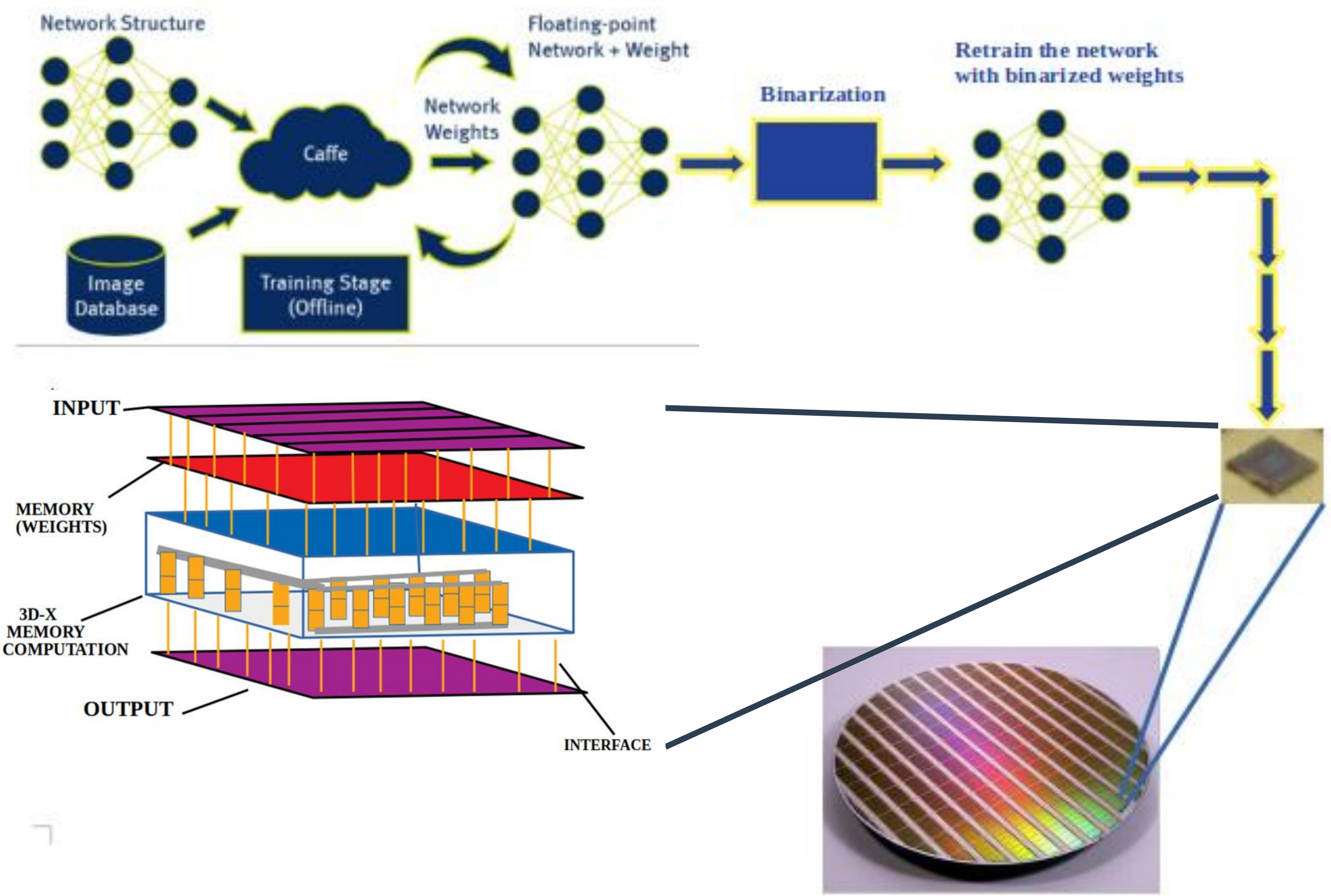


```
/test_bench_nn_1/x               32'b1111111111...   32'b000000...  32'b000100...  32'b001100...
/test_bench_nn_1/w               32'b1110001111...   32'b11100010001000100001...   32'b111000...
/test_bench_nn_1/w_layer_2       16'b1111101110...   16'b1011101110000011          16'b111110...
/test_bench_nn_1/neuron_top_o... 2'b01               2'b10                         2'b00
/test_bench_nn_1/Errors          TRUE
/test_bench_nn_1/ErrorCount      32'b0000000000...   32'b000000...  32'b000000...  32'b0000000000
```



All this in a single file/library

PRODUCT CONCEPTION


Network Structure
Image Database
Caffe
Training Stage (Offline)
Network Weights
Floating-point Network + Weight
Binarization
Retrain the network with binarized weights
INPUT
MEMORY (WEIGHTS)
3D-X MEMORY COMPUTATION
OUTPUT
INTERFACE