IL2230 HARDWARE ARCHITECTURE FOR DEEP LEARNING

Lab 1

Andreas Lezdins, Patrik Johansson, Chuwei Li, Jiayi Feng

Introduction

One of the fundamental building blocks of neural networks is the artificial neuron. The mathematical representation of the artificial neuron is a weighted sum of m inputs that is passed through a non-linear activation function, f. Figure 1 shows the concept of the artificial neuron.

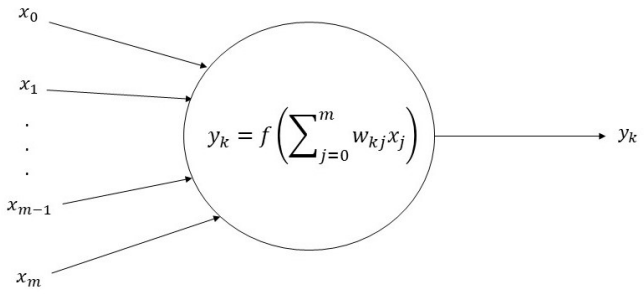


Figure 1: Basic concept on an artificial neuron.

When constructing an artificial neuron, the core computational functionality can be described by a Multiplier-Accumulator (MAC) unit. Figure 2 illustrates a MAC-unit which takes three values as input, lets assume a, b and c, and produces a result given by $y = (a \times b) + c$.

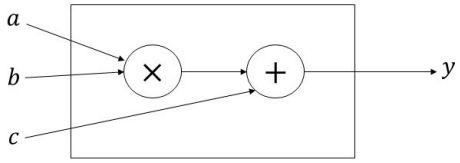


Figure 2: Computational concept of an MAC-unit.

Method

This report aims to design and evaluate an N-input artificial neuron using three different structures. The first implementation is a fully serial N-input structure. This is constructed using one single MAC-unit letting the control path be responsible for handling the repetitive operations. The second implementation is a fully parallel N-input structure connecting N MAC-units in parallel. The third implementation is a semi-parallel N-input structure. The structure is divided into K branches meaning the structure is composed of K MAC-units, in this report, we used $K = 2$. The three structures are presented in Figure 3.

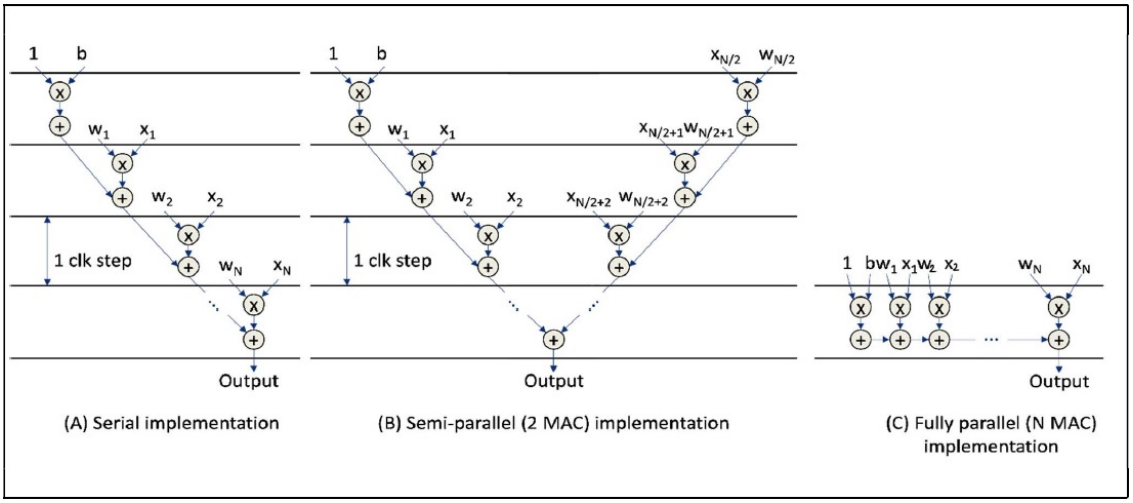


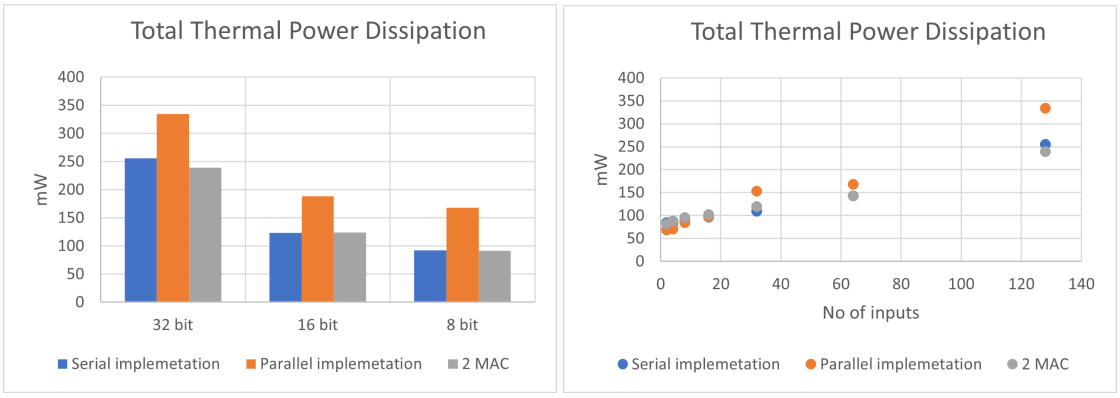
Figure 3: The three implemented structures.

Simulations of the structures was performed by varying $N = \{2,4,8,16,32,64,128\}$. The input and output data is set to 32 bit fixed point numbers where 12 bits were used for the integer part and 20 bits used for the fractional part. With $N = 128$ all structures were also simulated with 16 bits and 8 bits fixed-point numbers. With 6 bits and 3 bits as the integer part and 10 bits and 5 bits as the fractional part, respectively. The activation function used is the Rectifier Linear Unit (ReLU) function, given by $f(x) = max(0,x)$.

All simulations were performed in the Intel Quartus II software using Cyclon III FPGA with 9 bit built-in multiplier element. Information regarding the specific hardware used in the simulations is found in Appendix A. The implementations was simulated by connecting them to a dummy memory structure containing both the weights W and the inputs x. The dummy memory had a serial input and parallel output to the implementations tested.

Results

From the simulations, Figure 4a shows the power dissipation from the three structures when the length of the input and the output data varies. We see that increasing the length of the input and output data, increases the power dissipation. Figure 4b shows the power dissipation when varying N, with increasing N the power dissipation also increases.

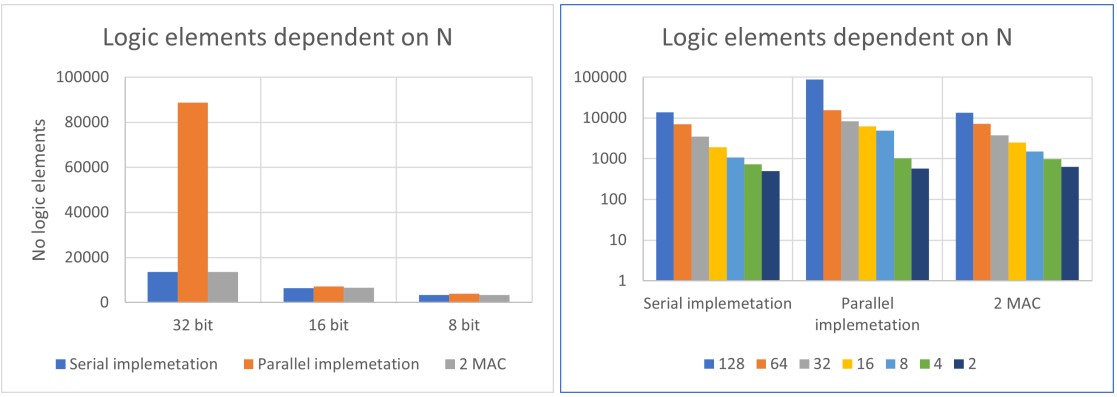


(a) Power dissipation when varying input data length.

(b) Power dissipation varying N with 32 bits input data.

Figure 4: Power dissipation from simulations.

Figure 5a shows the number of logic elements used with $N = 128$ and varying the input data length. We see that 32 bits of input and output data on the fully parallel structure use the most logical elements. Figure 5b shows the area when varying N, we see that for lower N the fewer logic elements is needed. In general, the fully parallel structure requires more logical elements than the fully serial and the partially serial structures.

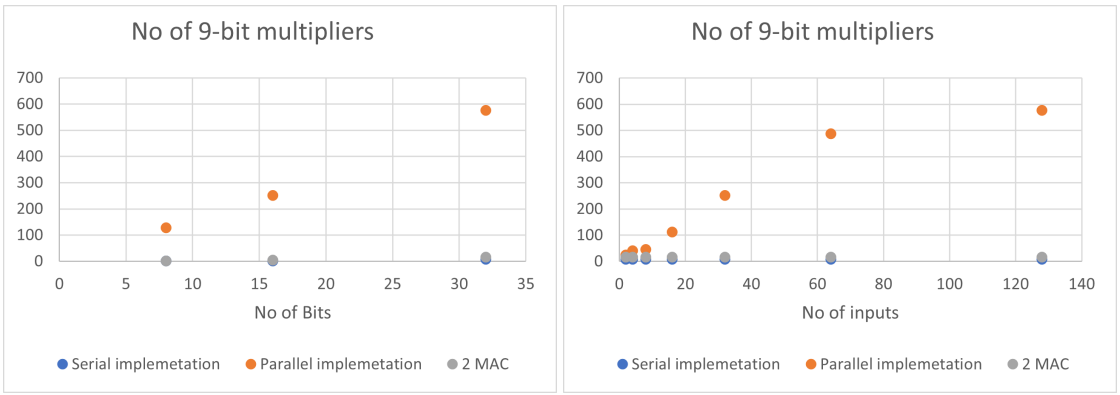


(a) Area with $N = 128$ and varying input data length.

(b) Area varying N with 32 bits input data.

Figure 5: Logic elements used when simulating the structures.

The number of multipliers used during simulations is shown in Figure 6. Figure 6a shows $N = 128$ and varying length of the input and output data, and Figure 6b shows varying N. We see that the number of multipliers used follows the same pattern as Figure 5 where larger N and larger input and output data requires more multipliers.

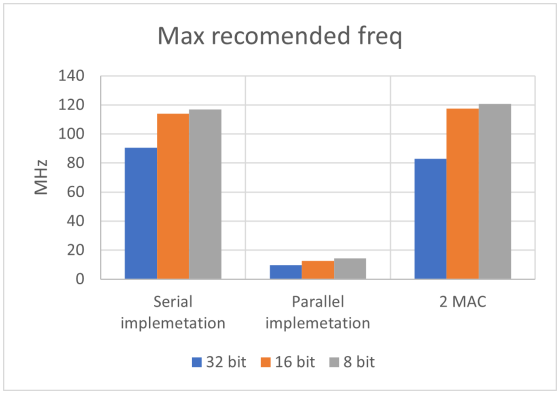


(a) Multipliers used with $N = 128$ and varying input data size.

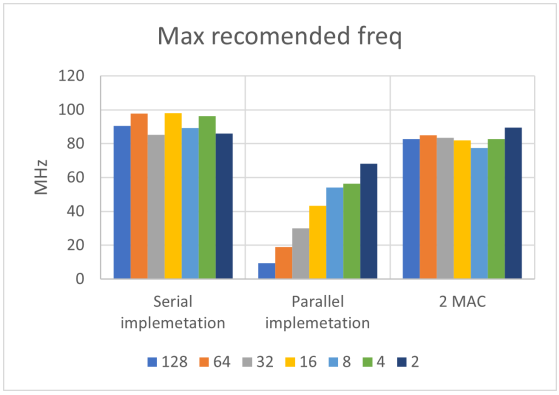
(b) Multipliers used varying N with 32 bits input data.

Figure 6: Multipliers used when simulating the structures.

The frequency of the three structures is shown in Figure 7, where Figure 7a shows the frequency with $N = 128$ and the length of the input and the output data is varied. We see that the parallel implementation has a substantially lower frequency than the two other structures, around six times lower. Figure 7b represents the frequency when N is varied, we can see that for lower N, the frequency used for the three structures is close, but as N increases, the lower the frequency gets for the fully parallel structure while the two other structures remain at the same frequency.



(a) Frequency used when $N = 128$ with varying input data.



(b) Frequency used varying N with 32 bits input data.

Figure 7: Frequency used when simulating the structures.

Conclusion

From the results, we can see that Figure 6 follows an almost linear pattern, which is to be expected. As the number of inputs increases the number of logical elements increases as well. With $N = 128$ we can see in Figure 6b that the number of multipliers used deviates somewhat from the linear behavior. This is because we ran the simulations using an FPGA, the Quartus II software made optimizations using logic elements to implement multipliers and thereby reducing the total number of multipliers when N increases. The same conclusion can be drawn from Figure 5, we clearly see that the number of logical element decreases when N decreases, which is to be expected.

The power dissipation also follows an almost liner pattern. Figure 4 shows that whit increasing N the power dissipation also increases. With $N = 128$ and varying the length of the input and output data we see that the fully parallel structure consumes more power than the other two structures. This is to be expected since the number of logical elements is much higher than the other structures.

The frequency displayed in Figure 7 shows that the fully parallel structure has a lower frequency than the other two. This is expected since the critical path is much longer than the other two. This is further strengthened in Figure 7b where for small N the number of MAC-entities used in the fully parallel structure is closer to the other two which has $N = 1$ and $N = 2$, here the frequency is close between all three structures. But as the number of MAC-units increases, the critical path increases, and the frequency lowers.

In terms of scalability, both the fully serial and the semi-parallel structures uses a fixed number of MAC-units, meaning they have endless scalability possibilities. The fully parallel structure scales the number of MAC-units with the number of inputs. Therefore there is a limit to the scalability of the structure. The advantage of the fully parallel structure is that all computations are completed within one clock cycle. This means the performance in terms of throughput scales with N. For higher N, the throughput of the fully parallel structure will be higher than the other two however, the need for more logical elements when N increases also means higher power dissipation and lower frequency for the fully parallel implementation compared to the other two implementations.

For further studies it is important to consider that the data that should supply the three structures needs to be transferred to the inputs. For a 32 bit 128 input neuron that produces data at a rate of 20 MHz this would equal approximately 9.5 Terabyte per second.

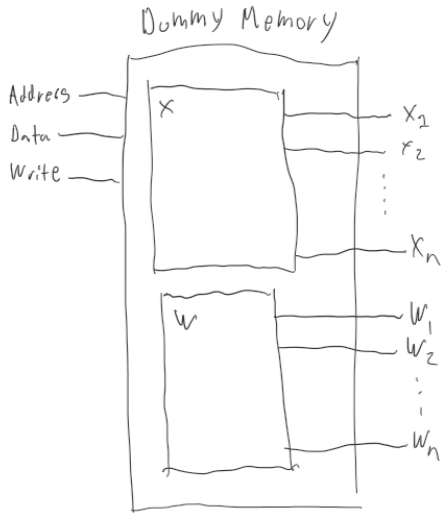
A Result data

Implemetation	N	Bits (fractional)	Clock target for the synthesis	Fmax (85c)	Total logic elements	Total combinational functions	Dedicated logic registers	Embedded Multiplier 9-bit elements	Family	Device	Total Thermal Power Dissipation	Core Dynamic Thermal Power Dissipation	Core Static Thermal Power Dissipation	I/O Thermal Power Dissipation
serial	128	32(20)	50Mhz	72.35 MHz	12,203 / 24,624 (50 %)	10,121 / 24,624 (41 %)	8,393 / 24,624 (34 %)	8 / 132 (6 %)	Cyclone III		166.10 mW	61.78 mW	82.53 mW	21.80 mW
serial	64	32(20)	50Mhz	76.0 MHz	6,308 / 10,320 (61 %)	5,251 / 10,320 (51 %)	4,264 / 10,320 (41 %)	8 / 46 (17 %)	Cyclone III		99.62 mW	31.91 mW	46.84 mW	20.87 mW
serial	32	32(20)	50Mhz	75.54 MHz	3,289 / 5,136 (64 %)	2,745 / 5,136 (53 %)	2,247 / 5,136 (44 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	86.45 mW	19.07 mW	46.48 mW	20.90 mW
serial	16	32(20)	50Mhz	80.58 MHz	1,843 / 5,136 (36 %)	1,555 / 5,136 (30 %)	1,190 / 5,136 (23 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	80.26 mW	13.03 mW	46.33 mW	20.91 mW
serial	8	32(20)	50Mhz	76.36 MHz	1,052 / 5,136 (20 %)	892 / 5,136 (17 %)	709 / 5,136 (14 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	76.30 mW	9.19 mW	46.23 mW	20.88 mW
serial	4	32(20)	50Mhz	81.1 MHz	720 / 5,136 (14 %)	624 / 5,136 (12 %)	420 / 5,136 (8 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	73.35 mW	6.32 mW	46.20 mW	20.83 mW
serial	2	32(20)	50Mhz	68.42 MHz	490 / 5,136 (10 %)	426 / 5,136 (8 %)	323 / 5,136 (6 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	73.47 mW	6.42 mW	46.18 mW	20.86 mW
serial	128	32(20)	100Mhz	13,681 / 24,624 (56 %)	10,121 / 24,624 (41 %)	8,393 / 24,624 (34 %)	8 / 132 (6 %)	Cyclone III	EP3C25F256C6	255.83 mW	146.43 mW	82.82 mW	26.58 mW
serial	64	32(20)	100Mhz	97.79 MHz	7,005 / 10,320 (68 %)	5,251 / 10,320 (51 %)	4,264 / 10,320 (41 %)	8 / 46 (17 %)	Cyclone III	EP3C10F256C6	142.65 mW	70.04 mW	46.94 mW	25.66 mW
serial	32	32(20)	100Mhz	85.32 MHz	3,460 / 5,136 (67 %)	2,745 / 5,136 (53 %)	2,247 / 5,136 (44 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	109.27 mW	36.96 mW	46.50 mW	25.80 mW
serial	16	32(20)	100Mhz	98.15 MHz	1,920 / 5,136 (37 %)	1,555 / 5,136 (30 %)	1,190 / 5,136 (23 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	96.78 mW	24.66 mW	46.35 mW	25.77 mW
serial	8	32(20)	100Mhz	89.37 MHz	1,085 / 5,136 (21 %)	892 / 5,136 (17 %)	709 / 5,136 (14 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	88.50 mW	16.53 mW	46.25 mW	25.73 mW
serial	4	32(20)	100Mhz	96.23 MHz	724 / 5,136 (14 %)	624 / 5,136 (12 %)	420 / 5,136 (8 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	84.79 mW	12.80 mW	46.21 mW	25.78 mW
serial	2	32(20)	100Mhz	86.01 MHz	497 / 5,136 (10 %)	426 / 5,136 (8 %)	323 / 5,136 (6 %)	8 / 46 (17 %)	Cyclone III	EP3C5F256C6	84.67 mW	12.70 mW	46.19 mW	25.78 mW
parallel	128	32(20)	100Mhz	9.53 MHz	88,765 / 119,088 (75 %)	87,385 / 119,088 (73 %)	8,288 / 119,088 (7 %)	576 / 576 (100 %)	Cyclone III	EP3C120F484C7	334.37 mW	202.04 mW	104.86 mW	27.47 mW
parallel	64	32(20)	100Mhz	18.88 MHz	15,759 / 81,264 (19 %)	13,898 / 81,264 (17 %)	2,308 / 81,264 (3 %)	488 / 488 (100 %)	Cyclone III	EP3C80F484C6	168.44 mW	46.79 mW	99.57 mW	22.09 mW
parallel	32	32(20)	100Mhz	30.05 MHz	8,260 / 39,600 (21 %)	7,336 / 39,600 (19 %)	2,144 / 39,600 (5 %)	252 / 252 (100 %)	Cyclone III	EP3C40F324C6	153.08 mW	37.73 mW	89.56 mW	25.80 mW
parallel	16	32(20)	100Mhz	43.48 MHz	6,168 / 15,408 (40 %)	5,823 / 15,408 (38 %)	1,120 / 15,408 (7 %)	112 / 112 (100 %)	Cyclone III	EP3C16F256C6	97.05 mW	21.60 mW	52.09 mW	23.36 mW
parallel	8	32(20)	100Mhz	54.03 MHz	4,864 / 10,320 (47 %)	4,710 / 10,320 (46 %)	608 / 10,320 (6 %)	46 / 46 (100 %)	Cyclone III	EP3C10F256C6	84.52 mW	17.15 mW	46.63 mW	20.74 mW
parallel	4	32(20)	100Mhz	56.33 MHz	1,024 / 5,136 (20 %)	896 / 5,136 (17 %)	320 / 5,136 (6 %)	40 / 46 (87 %)	Cyclone III	EP3C5F256C6	70.41 mW	6.76 mW	46.22 mW	17.43 mW
parallel	2	32(20)	100Mhz	68.27 MHz	580 / 5,136 (11 %)	516 / 5,136 (10 %)	192 / 5,136 (4 %)	24 / 46 (52 %)	Cyclone III	EP3C5F256C6	68.72 mW	4.52 mW	46.18 mW	18.02 mW
2 MAC	128	32(20)	50Mhz	82.78 MHz	13,565 / 24,624 (55 %)	10,665 / 24,624 (43 %)	8,488 / 24,624 (34 %)	16 / 132 (12 %)	Cyclone III	EP3C25F256C6	239.25 mW	129.97 mW	82.68 mW	26.60 mW
2 MAC	64	32(20)	50Mhz	85.08 MHz	7,127 / 10,320 (69 %)	5,716 / 10,320 (55 %)	4,359 / 10,320 (42 %)	16 / 46 (35 %)	Cyclone III	EP3C10F256C6	148.76 mW	71.08 mW	46.91 mW	25.78 mW
2 MAC	32	32(20)	50Mhz	83.4 MHz	3,788 / 10,320 (37 %)	3,159 / 10,320 (31 %)	2,310 / 10,320 (22 %)	16 / 46 (35 %)	Cyclone III	EP3C10F256C6	119.58 mW	47.30 mW	46.57 mW	25.71 mW
2 MAC	16	32(20)	50Mhz	81.88 MHz	2,503 / 5,136 (49 %)	2,164 / 5,136 (42 %)	1,317 / 5,136 (26 %)	16 / 46 (35 %)	Cyclone III	EP3C5F256C6	103.03 mW	30.94 mW	46.41 mW	25.69 mW
2 MAC	8	32(20)	50Mhz	77.53 MHz	1,495 / 5,136 (29 %)	1,274 / 5,136 (25 %)	836 / 5,136 (16 %)	16 / 46 (35 %)	Cyclone III	EP3C5F256C6	95.85 mW	23.78 mW	46.30 mW	25.76 mW
2 MAC	4	32(20)	50Mhz	82.62 MHz	977 / 5,136 (19 %)	842 / 5,136 (16 %)	547 / 5,136 (11 %)	16 / 46 (35 %)	Cyclone III	EP3C5F256C6	88.32 mW	16.39 mW	46.23 mW	25.69 mW
2 MAC	2	32(20)	50Mhz	89.57 MHz	632 / 5,136 (12 %)	570 / 5,136 (11 %)	291 / 5,136 (6 %)	16 / 46 (35 %)	Cyclone III	EP3C5F256C6	82.08 mW	10.12 mW	46.20 mW	25.76 mW
parallel	128	16(10)	50Mhz	9.53 MHz	88,765 / 119,088 (75 %)	87,385 / 119,088 (73 %)	8,288 / 119,088 (7 %)	576 / 576 (100 %)	Cyclone III	EP3C120F484C7	334.37 mW	202.04 mW	104.86 mW	27.47 mW
parallel	64	16(10)	50Mhz	12.71 MHz	7,191 / 39,600 (18 %)	7,129 / 39,600 (18 %)	2,128 / 39,600 (5 %)	252 / 252 (100 %)	Cyclone III	EP3C40F324C6	188.65 mW	76.03 mW	89.62 mW	23.00 mW
parallel	32	16(10)	50Mhz	14.32 MHz	3,863 / 81,264 (5 %)	3,567 / 81,264 (4 %)	1,048 / 81,264 (1 %)	128 / 488 (26 %)	Cyclone III	EP3C40F324C6	167.70 mW	50.40 mW	98.74 mW	18.57 mW
serial	128	32(20)	100Mhz	90.51 MHz	13,681 / 24,624 (56 %)	10,121 / 24,624 (41 %)	8,393 / 24,624 (34 %)	8 / 132 (6 %)	Cyclone III	EP3C25F256C6	255.83 mW	146.43 mW	82.82 mW	26.58 mW
serial	128	16(10)	100Mhz	114.03 MHz	6,489 / 10,320 (63 %)	5,166 / 10,320 (50 %)	4,201 / 10,320 (41 %)	2 / 46 (4 %)	Cyclone III	EP3C10F256C6	123.03 mW	58.06 mW	46.89 mW	18.09 mW
serial	128	8(5)	100Mhz	116.86 MHz	3,284 / 5,136 (64 %)	2,733 / 5,136 (53 %)	2,105 / 5,136 (41 %)	1 / 46 (2 %)	Cyclone III	EP3C5F256C6	92.35 mW	31.58 mW	46.51 mW	14.26 mW
2 MAC	128	32(20)	100Mhz	82.79 MHz	13,565 / 24,624 (55 %)	10,665 / 24,624 (43 %)	8,488 / 24,624 (34 %)	16 / 132 (12 %)	Cyclone III	EP3C25F256C6	239.25 mW	129.97 mW	82.68 mW	26.60 mW
2 MAC	128	16(10)	100Mhz	117.5 MHz	6,568 / 10,320 (64 %)	5,394 / 10,320 (52 %)	4,248 / 10,320 (41 %)	4 / 46 (9 %)	Cyclone III	EP3C10F256C6	124.10 mW	59.23 mW	46.87 mW	18.00 mW
2 MAC	128	8(5)	100Mhz	120.58 MHz	3,384 / 5,136 (66 %)	2,849 / 5,136 (55 %)	2,128 / 5,136 (41 %)	2 / 46 (4 %)	Cyclone III	EP3C5F256C6	91.48 mW	30.70 mW	46.50 mW	14.28 mW

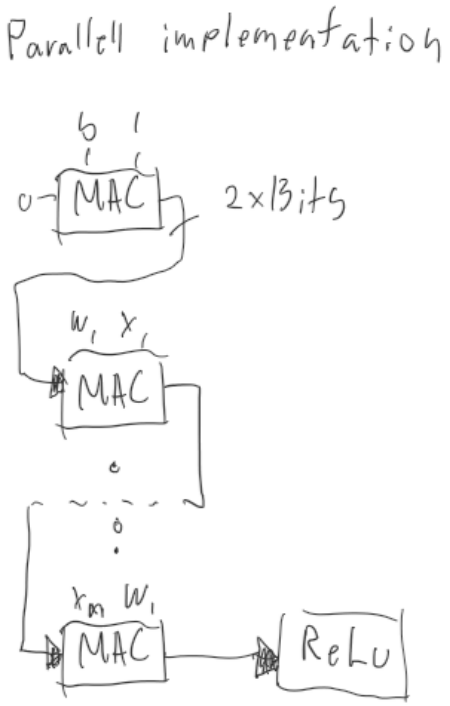
Figure 8: The raw data from the synthesis with Quartus II.

B Block Diagrams

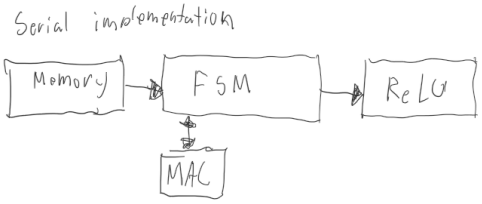
Block diagrams for the implementations, and simulations.



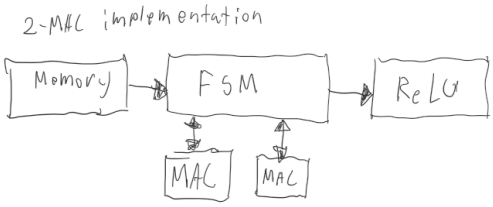
(a) Block diagram for the dummy memory.



(b) Block diagram for the parallel implementation.



(a) Block diagram for the serial implementation.



(b) Block diagram for the 2-Mac implementation.