A Report on

Low-Power VLSI Implementation of CNN on Accelerated 2D Systolic Array

Team RizzNet (CE)

Krish Mehta

Aryan Devrani

Anoushka Saraswat

Kumar Divij

ECE 284
University of California, San Diego (UCSD)
16th Dec. 2023

1. Motivation

2D Systolic Architecture enables modern high throughput AI/ML computation. We aim to implement an optimized 2D Systolic Array Design, with the goal of learning the ins and outs of its implementation, and various possible enhancements to the baseline architecture. The improvements can be broadly categorized into:

a) Parallelizing the Systolic Array Computation Stages
b) Reducing FIFO Depth from 64 to 16
c) Sparsity Aware Clock Gating
d) Scalable RTL Design for Multi-channel PEs
e) Additional Analysis/Verification

2. Parallelizing Systolic Array Computation

In the baseline model, all Kernel Loadings & MAC Operations are processed in a serial manner. In order to reduce the total number of execution cycles, these operations were pipelined to execute parallelly by modifying the control logic of the MAC Array Architecture:

2.1 Simultaneous Read/Write for Input & Output FIFOs

The L0 (IFIFO) and OFIFO can perform both Read & Write operations in the same cycle simultaneously, thus reducing the number of cycles taken to serially perform the two tasks.

2.2 Pipelining of MAC Load & Execute Operations

Instead of waiting for all Kernel Loadings to complete, when PE(x,y) is loading its weights, PE(x, y-2) starts its execution as it has already received its inputs. This increases the overall utilization of the MAC array as it stays idle for a much lesser time.

2.3 No Reset for L0 and OFIFO between Weight Change

This allows for new weights to start loading (from SRAM->L0->MAC) while the previous PSUMs are still being written to Memory.

Improvements:

1. Total Execution Cycles: **1973 -> 917 (54% reduction)**
2. Full MAC Utilization: **17% -> 28% (1.65X increase)**

3. Reducing L0 & OFIFO Depth from 64 to 16

After achieving this level of parallelism in the MAC computations, the FIFOs do not require a 64 depth. Hence, the FIFO depths were reduced to 16 by using a single 16:1 mux (instead of 4).

Improvements:

1. Number of Logical Elements: **17063 -> 7199 (67% reduction)**
2. Core Dynamic Power Consumption: **31.86 mW -> 18.68 mW (40% reduction)**
3. F_{max} for Slow 100C Model: **129.95 MHz -> 131.6 MHz**
4. TOP/s-W : **0.522 -> 0.900**
5. Total Execution time: **15.18 μs (vanilla) -> 7.06 μs (parallelized) -> 6.97 μs (parallelized+fifo_16)**

The following graph shows the flow of Pipelined MAC Operations across multiple iterations:

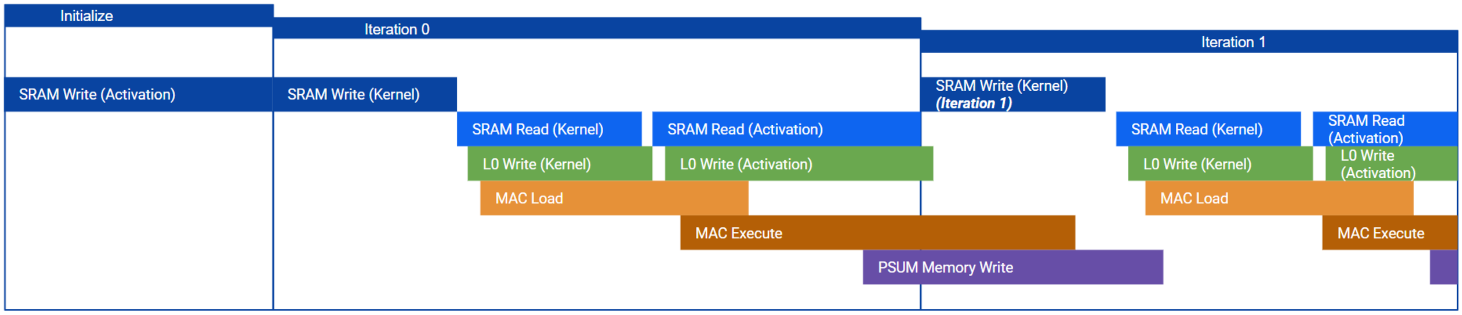
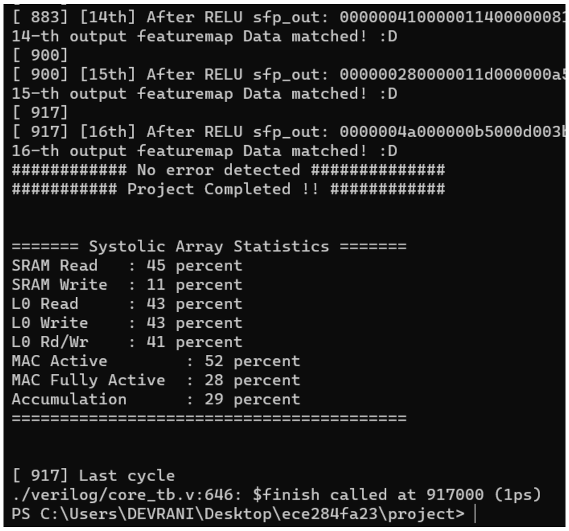


Fig.1 TImeline of Pipelined MAC Operations

In order to analyze and prove the enhancements achieved by these steps, we performed a Utilization Analysis for SRAM, L0 and MAC operations to observe the achieved parallelism:

Total Execution Time	1973 cycles (Serial) **917 cycles (Parallel)**
Stage	**Utilization (% of total execution time)**
MAC Active **(At least 1 PE Loading/Executing)**	34 % (Serial) **52 % (Parallel)**
MAC Fully Active **(Producing 8 PSUMs in a cycle)**	17 % (Serial) **28 % (Parallel)**
L0 FIFO Reading+Writing	0 % (Serial) **41 % (Parallel)**



```
[ 883] [14th] After RELU sfp_out: 0000004100000114000000081
14-th output featuremap Data matched! :D
[ 900]
[ 900] [15th] After RELU sfp_out: 000000280000011d000000a5
15-th output featuremap Data matched! :D
[ 917]
[ 917] [16th] After RELU sfp_out: 0000004a000000b5000d003b
16-th output featuremap Data matched! :D
############ No error detected ##############
########### Project Completed !! ############


======= Systolic Array Statistics =======
SRAM Read   : 45 percent
SRAM Write  : 11 percent
L0 Read     : 43 percent
L0 Write    : 43 percent
L0 Rd/Wr    : 41 percent
MAC Active         : 52 percent
MAC Fully Active   : 28 percent
Accumulation       : 29 percent
=========================================


[ 917] Last cycle
./verilog/core_tb.v:646: $finish called at 917000 (1ps)
PS C:\Users\DEVRANI\Desktop\ece284fa23\project>
```

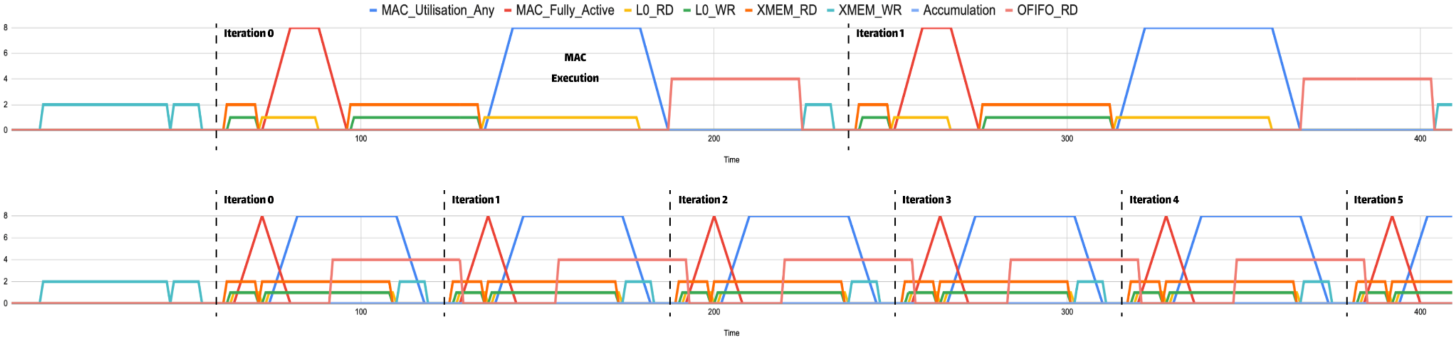


Fig.2(a) Statistics for achieved % Utilization across different stages; 2(b) Verilog output for Utilization; 2(c) Utilization Graphs for Baseline (serialized) and the Alpha (parallelized) models.

4. Sparsity Aware Clock Gating

In order to reduce Dynamic Power consumption for a core, we can reduce the amount of toggling operations that are made even when the inputs are sparse.

4.1 Structured Pruning

Structured pruning on the 4-bit Quantized VGGNet model allowed us to introduce high sparsity levels (>50%), and we were still able to recover >90% model accuracy on the sparse inputs with finetuning.

Model Type	Accuracy	Psum Error
4-bit Quantized VGGNet (with 8x8 Conv Layer)	92.07%	1.0455×10^{-7}
4-bit Quantized VGGNet (with 8x8 conv + 50% structured pruning)	91.72%	1.722×10^{-7}
4-bit Quantized VGGNet (with 8x8 conv + 70% structured pruning)	88.35%	1.3864×10^{-7}

Fig.3 Model Accuracies and Psum Error with quantization and pruning

4.2 Input-Channel Wise Clock Gating

- ➔ There are 8 clock lines going to each of the 8 mac_row modules, from there it passes to each of 8 mac_tile modules. Both the free running clock and gated clocks are sent to every mac_tile.
- ➔ Gated clocks are generated if ALL weights in a row are 0 (Zero-Condition).
- ➔ If Zero-Condition occurs, gated clock prevents new data latching and thereby multiplication calculations, by freezing the horizontal movement of data.

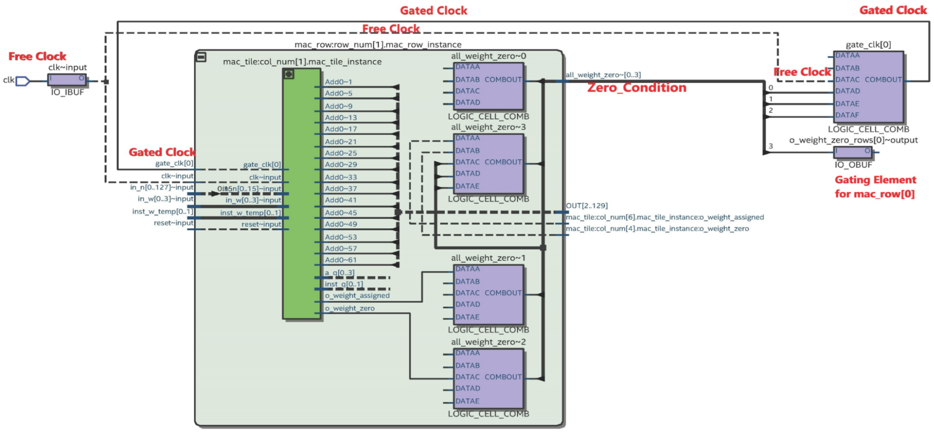


Fig.4 Structure of Input-Channel Wise Clock Gating

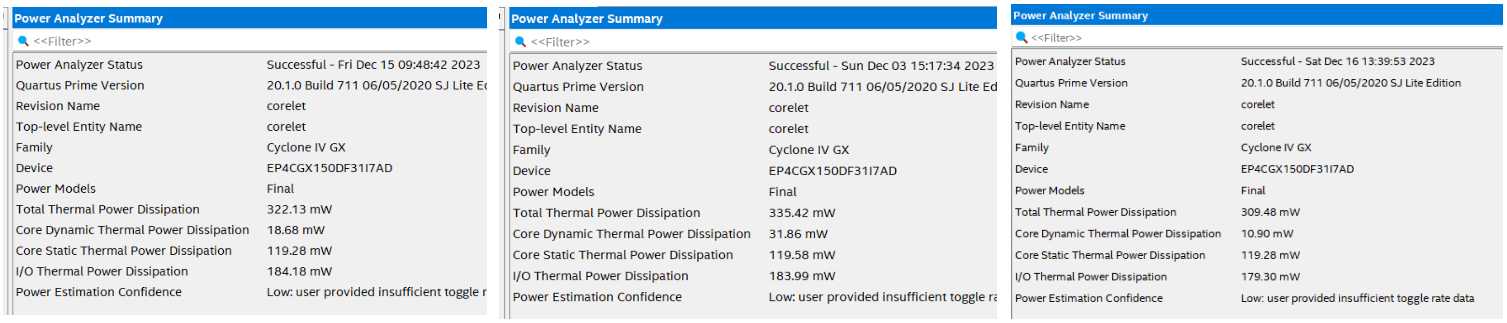
- ➔ Free running clock still ensures proper movement of partial sum data from north to south.

Improvements: Core Dynamic Power Consumption: **31.86 mW (Vanilla) -> 18.68mW (Parallelized + FIFO16) -> 10.90mW (Parallelized + FIFO16 + Clock Gated)**

5. Hardware Mapping, Timing & Power Analysis

We performed hardware mapping & analysis for 4 different models:

1. **VGGNet (Vanilla) -** MAC Operations are serialized.
2. **VGGNet Parallelized (Control Only) -** MAC Operations are pipelined, FIFO depth is 64
3. **VGGNet Parallelized (Control + FIFO Depth 64->16) -** Pipelined + FIFO Depth reduced to 16
4. **VGGNet Parallelized + Clock Gated -** Pipeline + FIFO Depth 16 + Input-channel Clock Gating

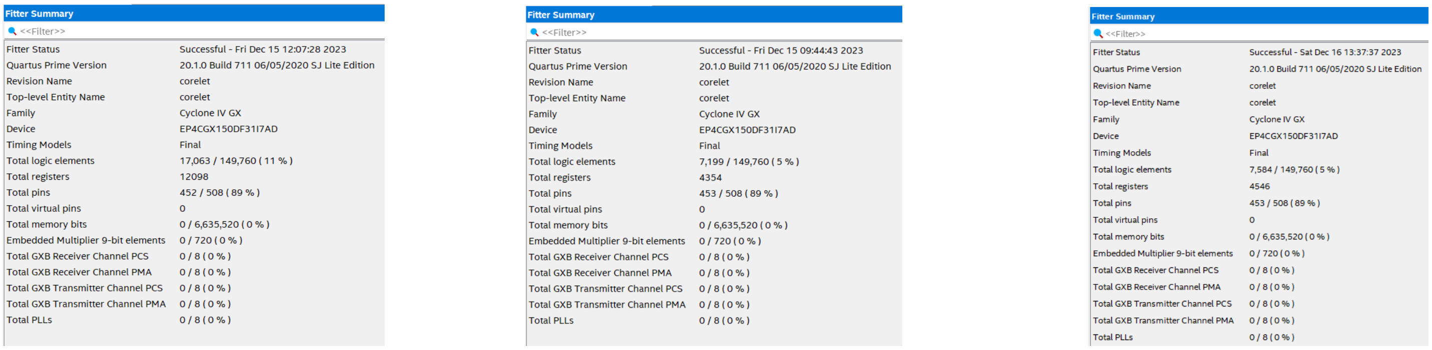


Power Analyzer Status	Successful - Fri Dec 15 09:48:42 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Ec
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Power Models	Final
Total Thermal Power Dissipation	322.13 mW
Core Dynamic Thermal Power Dissipation	18.68 mW
Core Static Thermal Power Dissipation	119.28 mW
I/O Thermal Power Dissipation	184.18 mW
Power Estimation Confidence	Low: user provided insufficient toggle r

Power Analyzer Status	Successful - Sun Dec 03 15:17:34 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Ed
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Power Models	Final
Total Thermal Power Dissipation	335.42 mW
Core Dynamic Thermal Power Dissipation	31.86 mW
Core Static Thermal Power Dissipation	119.58 mW
I/O Thermal Power Dissipation	183.99 mW
Power Estimation Confidence	Low: user provided insufficient toggle ra

Power Analyzer Status	Successful - Sat Dec 16 13:39:53 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Power Models	Final
Total Thermal Power Dissipation	309.48 mW
Core Dynamic Thermal Power Dissipation	10.90 mW
Core Static Thermal Power Dissipation	119.28 mW
I/O Thermal Power Dissipation	179.30 mW
Power Estimation Confidence	Low: user provided insufficient toggle rate data

Fig.5 Power Analysis for (a) Baseline (b) VGGNet Parallelized (Control+FIFO) & (c) VGGNet Parallelized + Clock Gated At their respective F_{max} frequencies



(a)

Fitter Summary	
Fitter Status	Successful - Fri Dec 15 12:07:28 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Timing Models	Final
Total logic elements	17,063 / 149,760 (11 %)
Total registers	12098
Total pins	452 / 508 (89 %)
Total virtual pins	0
Total memory bits	0 / 6,635,520 (0 %)
Embedded Multiplier 9-bit elements	0 / 720 (0 %)
Total GXB Receiver Channel PCS	0 / 8 (0 %)
Total GXB Receiver Channel PMA	0 / 8 (0 %)
Total GXB Transmitter Channel PCS	0 / 8 (0 %)
Total GXB Transmitter Channel PMA	0 / 8 (0 %)
Total PLLs	0 / 8 (0 %)

(b)

Fitter Summary	
Fitter Status	Successful - Fri Dec 15 09:44:43 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Timing Models	Final
Total logic elements	7,199 / 149,760 (5 %)
Total registers	4354
Total pins	453 / 508 (89 %)
Total virtual pins	0
Total memory bits	0 / 6,635,520 (0 %)
Embedded Multiplier 9-bit elements	0 / 720 (0 %)
Total GXB Receiver Channel PCS	0 / 8 (0 %)
Total GXB Receiver Channel PMA	0 / 8 (0 %)
Total GXB Transmitter Channel PCS	0 / 8 (0 %)
Total GXB Transmitter Channel PMA	0 / 8 (0 %)
Total PLLs	0 / 8 (0 %)

(c)

Fitter Summary	
Fitter Status	Successful - Sat Dec 16 13:37:37 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Timing Models	Final
Total logic elements	7,584 / 149,760 (5 %)
Total registers	4546
Total pins	453 / 508 (89 %)
Total virtual pins	0
Total memory bits	0 / 6,635,520 (0 %)
Embedded Multiplier 9-bit elements	0 / 720 (0 %)
Total GXB Receiver Channel PCS	0 / 8 (0 %)
Total GXB Receiver Channel PMA	0 / 8 (0 %)
Total GXB Transmitter Channel PCS	0 / 8 (0 %)
Total GXB Transmitter Channel PMA	0 / 8 (0 %)
Total PLLs	0 / 8 (0 %)

Fig.6 Synthesis Fitter Summary for (a) Baseline (b) Parallelized (Control+FIFO) & (c) VGGNet Parallelized + Clock Gated

(**NOTE**: Hardware mapping results are identical for Vanilla & Parallelized (Control Only), as only the control logic has changed, there is no change in RTL components. They differ in total execution cycles as summarized in the table below.)

Key Points:

-> Upon parallelizing the Control logic, total execution cycles are reduced by **54%**.
-> In the parallelized model, reducing FIFO depth to 16 allows higher f_{max} , reduces no. of logical elements by **67%**, reduces Dynamic power consumption by **40%**, and increases TOP/s-W by **72%**.
-> By introducing Clock Gating to the above model, the dynamic power further reduces by **40%**.
-> There is a tradeoff between execution time and power consumption upon introducing clock gating, as it reduces core dynamic power at the expense of a higher execution time.

(**NOTE:** Since Quartus prime uses randomized inputs for timing analysis, the true sparsity of inputs is not being utilized in these results. However, clock gating is still able to reduce the dynamic power consumption, which indicates that a **significant improvement** in power reduction will be achieved with the actual sparse inputs.)

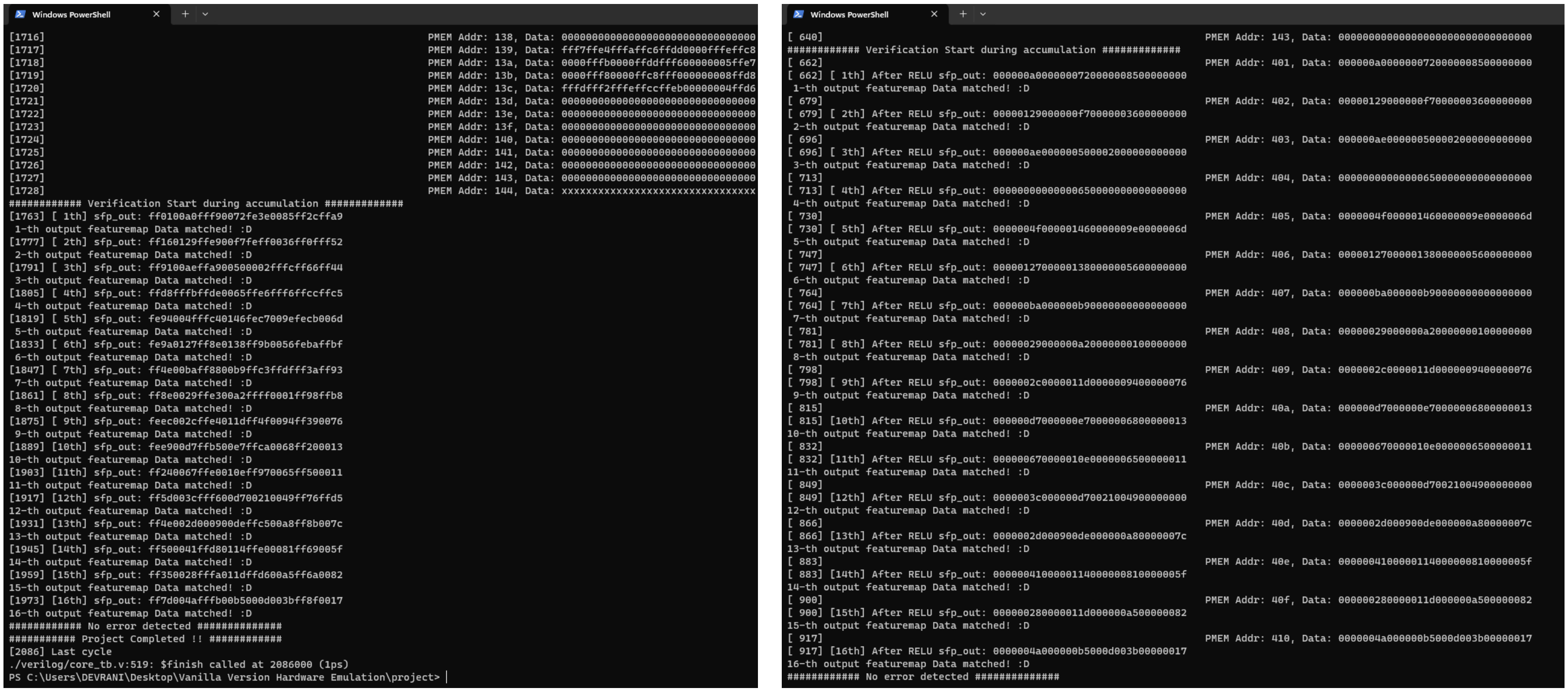
Parameter	VGGNet (Vanilla)	VGGNet Parallelized (Control only)	VGGNet Parallelized (Control + FIFO Depth 16)	VGGNet Parallelized + Clock Gated
f_{max} (Slow 100C Model)	129.95 MHz	129.95 MHz	131.6 MHz	80.78 MHz
Dynamic Power (at f_{max})	31.86 mW	31.86 mW	18.68 mW	10.90 mW
Power at matched freq. (80 Mhz)	19.76 mW	19.76 mW	11.41 mW	10.90 mW
TOP/s	0.0166	0.0166	0.0168	0.0103
TOP/s-W	0.522	0.522	0.900	0.945
Logical Elements	17063	17063	7199	7584
Total Execution Cycles	1973	917	917	917
Total Execution Time	15.18 µs	7.06 µs	6.97 µs	11.35 µs

🔴-> Worst 🟡-> Middle 🟢-> Best

6. Scalable RTL Design for Multichannel PEs

To increase the performance of our design further we modified the MAC PE to be able to compute multiplication and accumulation for 2 input channels simultaneously. We parameterized the TB, Core, Corelet, MAC Array/Row/Tile and the MAC PE such that the SRAM->L0->MAC Activation/Weight transfer path can scale its width, and the PE can perform the computation for multiple channels in the same cycle as per the parameter value. This allowed us to execute a 16 Input Channel x 8 Output Channel convolution in an 8x8 systolic array design.

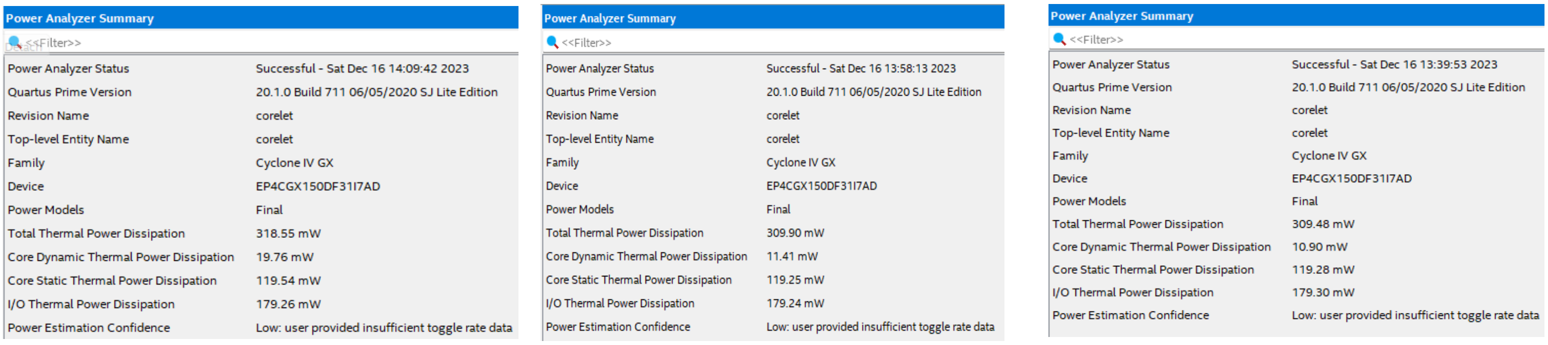
7. Additional Analysis/Verification



```
[1716]                                                       PMEM Addr: 138, Data: 00000000000000000000000000000000
[1717]                                                       PMEM Addr: 139, Data: fff7ffe4fffaffc6ffdd0000fffeffc8
[1718]                                                       PMEM Addr: 13a, Data: 0000fffb0000ffddfff600000005ffe7
[1719]                                                       PMEM Addr: 13b, Data: 0000fff80000ffc8fff000000008ffd8
[1720]                                                       PMEM Addr: 13c, Data: fffdfff2fffeffccffeb00000004ffd6
[1721]                                                       PMEM Addr: 13d, Data: 00000000000000000000000000000000
[1722]                                                       PMEM Addr: 13e, Data: 00000000000000000000000000000000
[1723]                                                       PMEM Addr: 13f, Data: 00000000000000000000000000000000
[1724]                                                       PMEM Addr: 140, Data: 00000000000000000000000000000000
[1725]                                                       PMEM Addr: 141, Data: 00000000000000000000000000000000
[1726]                                                       PMEM Addr: 142, Data: 00000000000000000000000000000000
[1727]                                                       PMEM Addr: 143, Data: 00000000000000000000000000000000
[1728]                                                       PMEM Addr: 144, Data: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
############ Verification Start during accumulation #############
[1763] [ 1th] sfp_out: ff0100a0fff90072fe3e0085ff2cffa9
 1-th output featuremap Data matched! :D
[1777] [ 2th] sfp_out: ff160129ffe900f7feff0036ff0fff52
 2-th output featuremap Data matched! :D
[1791] [ 3th] sfp_out: ff9100aeffa900500002fffcff66ff44
 3-th output featuremap Data matched! :D
[1805] [ 4th] sfp_out: ffd8fffbffde0065ffe6fff6ffccffc5
 4-th output featuremap Data matched! :D
[1819] [ 5th] sfp_out: fe94004fffc40146fec7009efecb006d
 5-th output featuremap Data matched! :D
[1833] [ 6th] sfp_out: fe9a0127ff8e0138ff9b0056febaffbf
 6-th output featuremap Data matched! :D
[1847] [ 7th] sfp_out: ff4e00baff8800b9ffc3ffdfff3aff93
 7-th output featuremap Data matched! :D
[1861] [ 8th] sfp_out: ff8e0029ffe300a2ffff0001ff98ffb8
 8-th output featuremap Data matched! :D
[1875] [ 9th] sfp_out: feec002cffe4011dff4f0094ff390076
 9-th output featuremap Data matched! :D
[1889] [10th] sfp_out: fee900d7ffb500e7ffca0068ff200013
10-th output featuremap Data matched! :D
[1903] [11th] sfp_out: ff240067ffe0010eff970065ff500011
11-th output featuremap Data matched! :D
[1917] [12th] sfp_out: ff5d003cfff600d700210049ff76ffd5
12-th output featuremap Data matched! :D
[1931] [13th] sfp_out: ff4e002d000900deffc500a8ff8b007c
13-th output featuremap Data matched! :D
[1945] [14th] sfp_out: ff500041ffd80114ffe00081ff69005f
14-th output featuremap Data matched! :D
[1959] [15th] sfp_out: ff350028fffa011dffd600a5ff6a0082
15-th output featuremap Data matched! :D
[1973] [16th] sfp_out: ff7d004afffb00b5000d003bff8f0017
16-th output featuremap Data matched! :D
############ No error detected ##############
########### Project Completed !! ############
[2086] Last cycle
./verilog/core_tb.v:519: $finish called at 2086000 (1ps)
PS C:\Users\DEVRANI\Desktop\Vanilla Version Hardware Emulation\project>
```

```
[ 640]                                                        PMEM Addr: 143, Data: 00000000000000000000000000000000
############ Verification Start during accumulation #############
[ 662]                                                        PMEM Addr: 401, Data: 000000a0000000720000008500000000
[ 662] [ 1th] After RELU sfp_out: 000000a0000000720000008500000000
 1-th output featuremap Data matched! :D
[ 679]                                                        PMEM Addr: 402, Data: 00000129000000f70000003600000000
[ 679] [ 2th] After RELU sfp_out: 00000129000000f70000003600000000
 2-th output featuremap Data matched! :D
[ 696]                                                        PMEM Addr: 403, Data: 000000ae000000500002000000000000
[ 696] [ 3th] After RELU sfp_out: 000000ae000000500002000000000000
 3-th output featuremap Data matched! :D
[ 713]                                                        PMEM Addr: 404, Data: 00000000000000650000000000000000
[ 713] [ 4th] After RELU sfp_out: 00000000000000650000000000000000
 4-th output featuremap Data matched! :D
[ 730]                                                        PMEM Addr: 405, Data: 0000004f000001460000009e0000006d
[ 730] [ 5th] After RELU sfp_out: 0000004f000001460000009e0000006d
 5-th output featuremap Data matched! :D
[ 747]                                                        PMEM Addr: 406, Data: 00000127000001380000005600000000
[ 747] [ 6th] After RELU sfp_out: 00000127000001380000005600000000
 6-th output featuremap Data matched! :D
[ 764]                                                        PMEM Addr: 407, Data: 000000ba000000b90000000000000000
[ 764] [ 7th] After RELU sfp_out: 000000ba000000b90000000000000000
 7-th output featuremap Data matched! :D
[ 781]                                                        PMEM Addr: 408, Data: 00000029000000a20000000100000000
[ 781] [ 8th] After RELU sfp_out: 00000029000000a20000000100000000
 8-th output featuremap Data matched! :D
[ 798]                                                        PMEM Addr: 409, Data: 0000002c0000011d0000009400000076
[ 798] [ 9th] After RELU sfp_out: 0000002c0000011d0000009400000076
 9-th output featuremap Data matched! :D
[ 815]                                                        PMEM Addr: 40a, Data: 000000d7000000e70000006800000013
[ 815] [10th] After RELU sfp_out: 000000d7000000e70000006800000013
10-th output featuremap Data matched! :D
[ 832]                                                        PMEM Addr: 40b, Data: 000000670000010e0000006500000011
[ 832] [11th] After RELU sfp_out: 000000670000010e0000006500000011
11-th output featuremap Data matched! :D
[ 849]                                                        PMEM Addr: 40c, Data: 0000003c000000d70021004900000000
[ 849] [12th] After RELU sfp_out: 0000003c000000d70021004900000000
12-th output featuremap Data matched! :D
[ 866]                                                        PMEM Addr: 40d, Data: 0000002d000900de000000a80000007c
[ 866] [13th] After RELU sfp_out: 0000002d000900de000000a80000007c
13-th output featuremap Data matched! :D
[ 883]                                                        PMEM Addr: 40e, Data: 00000041000001140000008100000005f
[ 883] [14th] After RELU sfp_out: 00000041000001140000008100000005f
14-th output featuremap Data matched! :D
[ 900]                                                        PMEM Addr: 40f, Data: 000000280000011d000000a500000082
[ 900] [15th] After RELU sfp_out: 000000280000011d000000a500000082
15-th output featuremap Data matched! :D
[ 917]                                                        PMEM Addr: 410, Data: 0000004a000000b5000d003b00000017
[ 917] [16th] After RELU sfp_out: 0000004a000000b5000d003b00000017
16-th output featuremap Data matched! :D
############ No error detected ##############
```

Fig.7 (a) Data Match output for Vanilla Model in 1973 cycles; (b) Data Match output for Parallelized Model in 917 cycles



Power Analyzer Status	Successful - Sat Dec 16 14:09:42 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Power Models	Final
Total Thermal Power Dissipation	318.55 mW
Core Dynamic Thermal Power Dissipation	19.76 mW
Core Static Thermal Power Dissipation	119.54 mW
I/O Thermal Power Dissipation	179.26 mW
Power Estimation Confidence	Low: user provided insufficient toggle rate data

Power Analyzer Status	Successful - Sat Dec 16 13:58:13 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Power Models	Final
Total Thermal Power Dissipation	309.90 mW
Core Dynamic Thermal Power Dissipation	11.41 mW
Core Static Thermal Power Dissipation	119.25 mW
I/O Thermal Power Dissipation	179.24 mW
Power Estimation Confidence	Low: user provided insufficient toggle rate data

Power Analyzer Status	Successful - Sat Dec 16 13:39:53 2023
Quartus Prime Version	20.1.0 Build 711 06/05/2020 SJ Lite Edition
Revision Name	corelet
Top-level Entity Name	corelet
Family	Cyclone IV GX
Device	EP4CGX150DF31I7AD
Power Models	Final
Total Thermal Power Dissipation	309.48 mW
Core Dynamic Thermal Power Dissipation	10.90 mW
Core Static Thermal Power Dissipation	119.28 mW
I/O Thermal Power Dissipation	179.30 mW
Power Estimation Confidence	Low: user provided insufficient toggle rate data

Fig.8 Power Analysis at common F=80 MHz for (a) Vanilla; (b) Parallelized(control+fifo16) & (c) Parallelized + Clock Gated

-> Comparing Dynamic power of the models at their respective F_{max} can be misleading as you would always expect a lower power at a lower freq. of operation ($P=CV^2Af$). Hence, analyzing them at a common frequency of f=80 MHz confirms that clock gating actually reduces dynamic power consumption over other models.

-> The tradeoff between power consumption & execution time depends on what the designer wants to optimize. However, if we look at the TOP/s-W metric which combines both Power and Time, the Clock Gated model clearly outperforms the rest.

END