Penerbit
Akademia Baru

Journal of Advanced Research in
Computing and Applications

Journal homepage: www.akademiabaru.com/arca.html
ISSN: 2462-1927

Low Power Integrated Circuit Design of Extreme Learning Machine using Power Gating Methodology

Chung Siong Shim[1], Chia Yee Ooi[1,*], Giap Seng Teoh[2]

[1] Embedded System iKohza, Department of Electronic Systems Engineering, Malaysia-Japan International Institute of Technology Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
[2] SkyeChip Pte Ltd, Bayan Lepas, Penang, Malaysia

ARTICLE INFO

Article history:
Received 11 February 2023
Received in revised form 20 April 2023
Accepted 19 May 2023
Available online 10 June 2023

ABSTRACT

With the current trend of individuals seeking more advanced and intelligent devices in their daily lives, artificial intelligence (AI) chips have become the primary focus for numerous industries. To create smarter AI chips, more computing power is often required, which can lead to an increase in power consumption. This paper describes a low power design of the Extreme Learning Machine (ELM) using single voltage power gating methodology to reduce power dissipation without affecting ELM's functionality. This is realized by applying power gating and state retention techniques in logic synthesis and physical design of ELM. First, a register transfer level (RTL) ELM was designed in Verilog based on the C code of ELM. Then, its functionality was verified by the testbench using VCS. Subsequently, Design Compiler (DC) tool, which was set-up with the inclusion of libraries containing power gating components, was used to synthesize the RTL into a netlist. The hotspot module was identified by compiling the design along with testbench, and timing constraints. The hotspot is where the power gating was implemented after it was described in Unified Power Format (UPF). Then, power analysis was done in PrimeTime tool. By switching power switch on and off, the total power was decreased from 0.1127V to 0.0784V which is 30.43% reduction. In physical design of ELM, power switches are synthesized in array style. Other physical layout design flow such as floorplan, power network synthesis, placement, and clock tree synthesis (CTS) were also performed.

Keywords:
Artificial intelligence, Extreme Learning Machine, Power Gating Methodology

1. Introduction

Currently, in the semiconductor market, there is a widespread shift in focus towards deep learning chips or AI chips by various industries. Initially, the two prominent chip design software companies, Synopsys and Cadence, announced the integration of machine learning into their software suites. However, in 2022, Google entered the scene and introduced PRIME, a deep-learning

[*] *Corresponding author*
E-mail address: ooichiayee@utm.my (Chia Yee Ooi)

approach for generating AI chip architecture [1]. To develop a more intelligent AI, machine learning algorithms necessitate larger models and extensive training sets, resulting in a substantial demand for power during data training and testing. The reliability of a system is compromised and the cost of implementing cooling systems for the chip increases due to excessive power consumption. Given the complexity of the circuit, which comprises multiple modules, it is wasteful to dissipate power in unused modules, so it is more rational to put them in an inactive mode when not in use. This paper demonstrates a low-power design of Extreme Learning Machine (ELM) with single voltage power gating methodology to reduce power dissipation without affecting ELM's functionality. The paper outline is as follows. Section 2 describes the basic components in ELM and low power techniques used in this work. Section 3 elaborates the low power design methodology and Section 4 details the experimental result. Section 5 concludes the work.

2. Extreme Learning Machine (ELM) and Low Power Gating

Single Hidden Layer Feedforward Neural Network (SLFN) is a subset of Feedforward Neural Network (FNN) with only one hidden layer. SLFN is widely used in Extreme Learning mainly for regression and classification [2]. A generalized SLFN has L hidden neurons and a set of N arbitrary distinct samples $(x_j, t_j); j = 1 ,\dots, N$, where $x_j = [x_{j1}, x_{j2}, \dots x_{n1}]^T \epsilon R^n$ are the input vectors and $t_j = [t_{j1}, t_{j2}, \dots t_{jm}]^T \epsilon R^m$ are the output vectors. The output vectors function of ELM, $y_j = [y_{j1}, y_{j2}, \dots y_{jm}]^T$, is formulated as Eq. (1).

$$t_j = y_j = \sum_{i=1}^{L} \beta_i h_i(x_j) = h(x_j)\beta \quad \mathrm{j} = 1, \dots, \mathrm{N} \tag{1}$$

where $\beta_i = [\beta_{i1}, \beta_{i2}, \dots \beta_{im}]^T \epsilon R^m$, i = 1 ,..., L, is the output weight vector connecting the i^{th} hidden layer neurons to output layer neurons, and $h(x_j) = [h_1(x_j), h_2(x_j), \dots h_L(x_j)]$ is the output (row) vector of the hidden layer with respect to input x_j , whereas $h_i(x_j)$ is the output function of the i^{th} hidden neuron. The $h_i(x_j)$ are not unique since different activation functions can be used in hidden neurons which is expressed as Eq. (2)

$$h_i(x_j) = G(a_i, b_i, x_j) \tag{2}$$

where $G(a_i, b_i, x_j)$ is an activation function that fulfils ELM universal approximation. $a_j = [a_{j1}, a_{j2}, \dots a_{n1}]^T \epsilon R^n$, $i = 1 ,\dots, N$, is the input weight vector connecting the input layer neurons to the i^{th} hidden neurons, and $b_i \epsilon R$ is the threshold of ith hidden neuron [3].

Power gating, also known as MTCMOS (Multi-threshold CMOS) is one of the most successful techniques to reduce standby leakage created when System-on-Chips (SoC) is developed below 90nm. Power gating is a process of shutting off a logic cluster when it is not in use, hence minimising run-time leakage power. This is accomplished by adding one or more MOSFETs to the supply network that are controlled by sleep signals. Power gating can be implemented in two ways which is header switch (PMOS transistor) and footer switch (NMOS transistor) as illustrated in Figure 1. Header switch is placed between the power supply and power-gated supply (virtual supply). Footer switch is place between the ground and power-gated ground (virtual ground) [4].

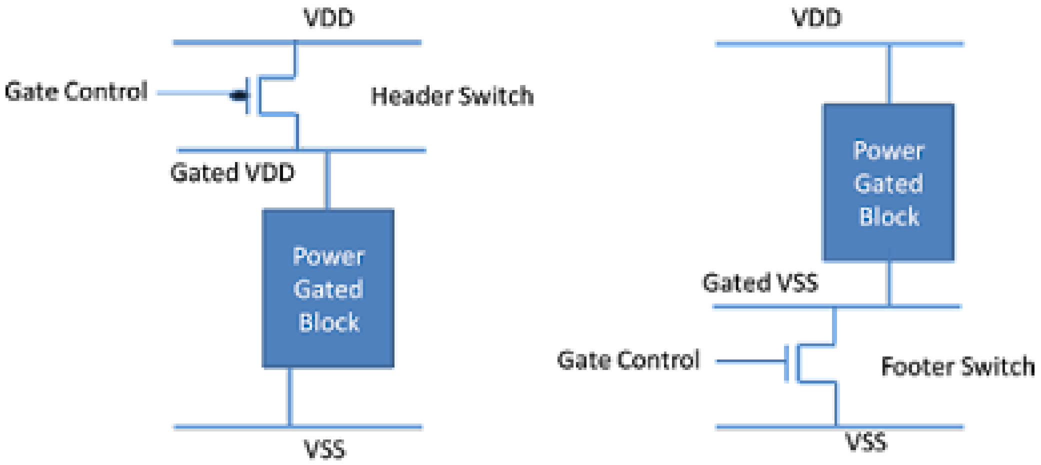


Fig. 1. Power gating of header switch and footer switch

3. Methodology

In this research, the flow of developing low power Extreme Learning Machine (ELM) in Very Large-scale Integration (VLSI) is shown as Figure 2. The tools applied in this project are Synopsys tools running on CentOS in VMware.

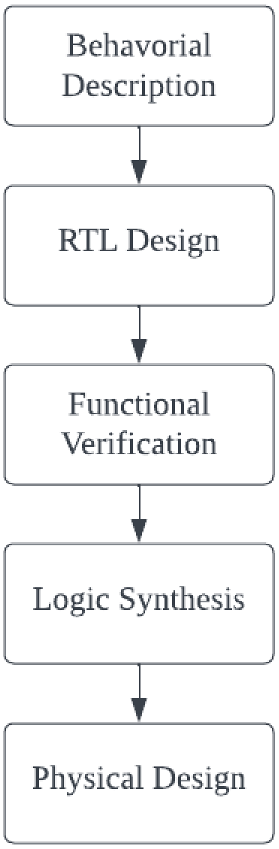


Fig. 2. Power gating of header switch and footer switch

Since the behavorial description of ELM is done previously in C Code and has been tested for correct functionality, deriving its Hardware Description Language (HDL) is the first task in this project. In this project, Verilog code will be used as the HDL to design ELM in Register Transfer Level (RTL). In the hidden layer of ELM neural network, it contains 20 hidden neurons and each hidden neuron takes 8 inputs to perform multiply-accumulate (MAC) and added with bias. The result from the MAC and summation of bias is used in activation function of ELM. The activation function used by ELM is Sigmoid function which is formulated as Eq. (3). The output neuron of ELM which also perform MAC function that takes 20 values from the outputs of 20 hidden neurons. The power controller is designed to control and shut down hidden layer when the block is not in used at the time. The signal of special cells such as retention cells, isolation cells and power switch.

$$G\left(a_i, b_i, x_j\right) = \frac{1}{1 + e^{-(a_i x_j + b_i)}} \tag{3}$$

Before logic synthesis, the Verilog code needs to be tested to ensure the functionality is the same as behavorial description in C code. A testbench is developed to provide input patterns to RTL design of ELM and then to compare the ELM outputs with the expected outputs. Along with the Verilog code of ELM, the testbench is compiled and simulated by Verilog Compiler Simulator (VCS). After compilation the result is showed based on the input inserted in testbench. In testbench, the output of C code is read from the text file and compare the result after data processed in RTL and print the result whether it is correct or wrong. Apart from verifying the RTL design of ELM, switching activities that take place in the design can be captured by setting the toggle region inside the testbench code. The output file of the switching activities is Switching Activity Interchange Format (SAIF) and it gives information of toggle rate of signal at each net.

In logic synthesis, Design Compiler (DC) tool from Synopsys is used to compile the design into a netlist format. Figure 3.7 shows the process and some important command lines applied in logic synthesis flow. It takes input files of Register Transfer Level (RTL) Verilog file, Unified Power Format (UPF) file, Switching Activity Interchange Format (SAIF) file and constraint files to compile the design. Logic synthesis flow is run twice in this project. The first run of the logic synthesis includes only RTL design, SAIF file and timing constraints. After first compilation, the power information of each block can be shown thus identify the hotspot block and create UPF file for implementation of power gating. The second compilation includes UPF file inside the flow to implement power gating technique into the design. UPF is a standard format used in the field of VLSI design to describe the power intent of a digital IC. Description of power-related information such as power domains, power supply connections, power states and declaration of special cells can be done in UPF.

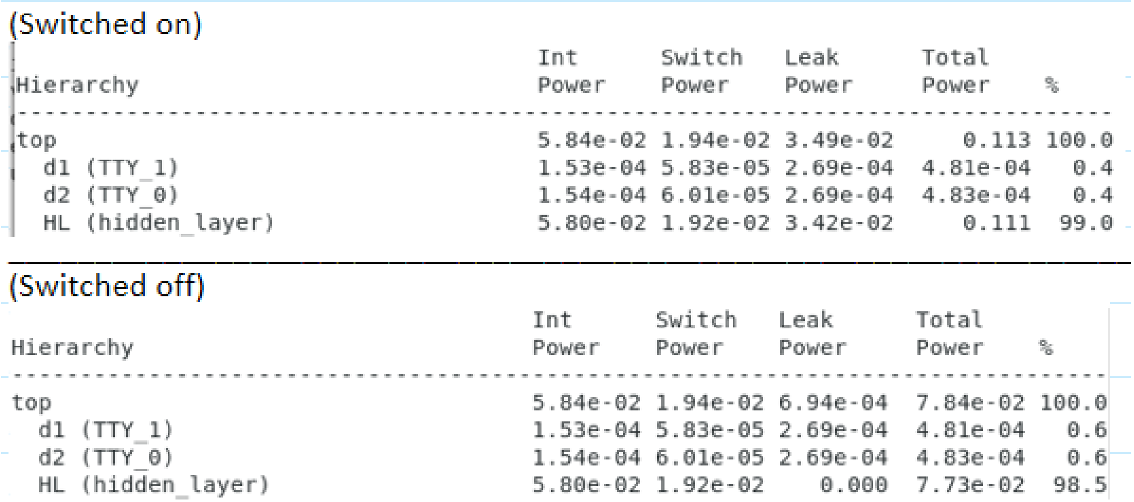
PrimeTime tool is used to provide power analysis capabilities by estimating the power consumption of the design based on the timing information. It allows designers to analyze power consumption at different levels, including individual cells, blocks, and the entire design. In this project, PrimeTime is used to observe the power decreased by power gating by comparing the power consumption between ELM design with and without power gating.

In physical design of ELM, floorplan stage defines the approximate locations and sizes of the design's major building blocks, power and ground distribution, and overall chip dimensions. Voltage area is also defined at this stage to match the power intent specified for each power domain during logic synthesis stage. All the cells that are used in the power domain will be placed in the same voltage area and the remaining will be put outside of voltage area using fast placement method. Power switches are synthesized right after fast placement of cells. Power switches are created in array style for always on characteristic. Next is power network synthesis which is Template-based Power Network Synthesis (TPNS) is used in this project. By using TPNS method, power rings, straps and rails are created for distributing power to every macro, standard cells, and all other cells are present in the design. The last two major steps in physical design of ELM are placement and clock tree synthesis (CTS). The placement tool determines the precise placement of each standard cell on the core which is set at between Metal Layer 4 to Metal Layer 9. CTS is used to evenly distribute the clock signal to all sequential components in a VLSI design by using command clock opt.

4. Results

The first compilation is done to determine the hotspot in ELM design while the second compilation of DC is a complete compilation of ELM augmented with exists of power gating. The second compilation takes input files including RTL design, UPF file, SAIF file, timing constraint and power constraint file. With the timing requirement is met with positive slack after compilation, the final verification is to confirm that special cells are created in the design such as isolation cell and retention cell. The verification is done by reporting all the special cells in DC.

In PrimeTime, power consumption of ELM circuit during switch on and that during switch off for hidden layer block are compared. Figure 3 shows the comparison of power consumption between the hidden layer block that is switched on and off.



```
(Switched on)
                                      Int      Switch   Leak     Total
Hierarchy                             Power    Power    Power    Power    %
---------------------------------------------------------------------------
top                                   5.84e-02 1.94e-02 3.49e-02    0.113 100.0
  d1 (TTY_1)                          1.53e-04 5.83e-05 2.69e-04 4.81e-04   0.4
  d2 (TTY_0)                          1.54e-04 6.01e-05 2.69e-04 4.83e-04   0.4
  HL (hidden_layer)                   5.80e-02 1.92e-02 3.42e-02    0.111  99.0
```

```
(Switched off)
                                      Int      Switch   Leak     Total
Hierarchy                             Power    Power    Power    Power    %
---------------------------------------------------------------------------
top                                   5.84e-02 1.94e-02 6.94e-04 7.84e-02 100.0
  d1 (TTY_1)                          1.53e-04 5.83e-05 2.69e-04 4.81e-04   0.6
  d2 (TTY_0)                          1.54e-04 6.01e-05 2.69e-04 4.83e-04   0.6
  HL (hidden_layer)                   5.80e-02 1.92e-02    0.000 7.73e-02  98.5
```

Fig. 3. Comparison of power consumption between the hidden layer block that is switched on and off

The leakage power of hidden layer block has been reduced to 0W as power switch shuts down the block and causes no leakage power to exist in the block. The main objective of power gating is to reduce the leakage power of the circuit. Thus, the power gating in ELM has successfully performed its function to decrease leakage power. The percentage of total power saving for power gating is calculated using the equation formulated in Eq. (4).

$$Total\ Power\ Saving = \frac{(Power\ Switch_{ON} - Power\ Switch_{OFF})}{Power\ Switch_{ON}} \times 100\% \tag{4}$$

$$\frac{0.1127 - 0.0784}{0.1127} \times 100\% = 30.43\%$$

By following proper settings and commands for ELM in physical design, ELM has successfully created floorplan with 0.2 core utilization and voltage area, PD_HL. Power switches are synthesized in array style after fast placement of standard cells. Power rings and power straps are also created using TPNS method. For vertical power rings of VDD_HL, VDD and VSS were created at Metal Layer M2 and Metal Layer M3 was used for horizontal. Since the power straps were created vertically, it is created at Metal Layer M3 except for VDD_HL power straps which was created at Metal Layer M4 to avoid congestion. After preroute standard cells which create power rails between Metal Layer M1 to Metal Layer M4, one of the major steps in physical design, optimized placement is executed and gives no DRC violations for the design. The physical design of ELM ended at CTS stage with no DRC violations. The power information in CTS is similar to power information in routing stage. Figure 4 shows the layout view of ELM in CTS.

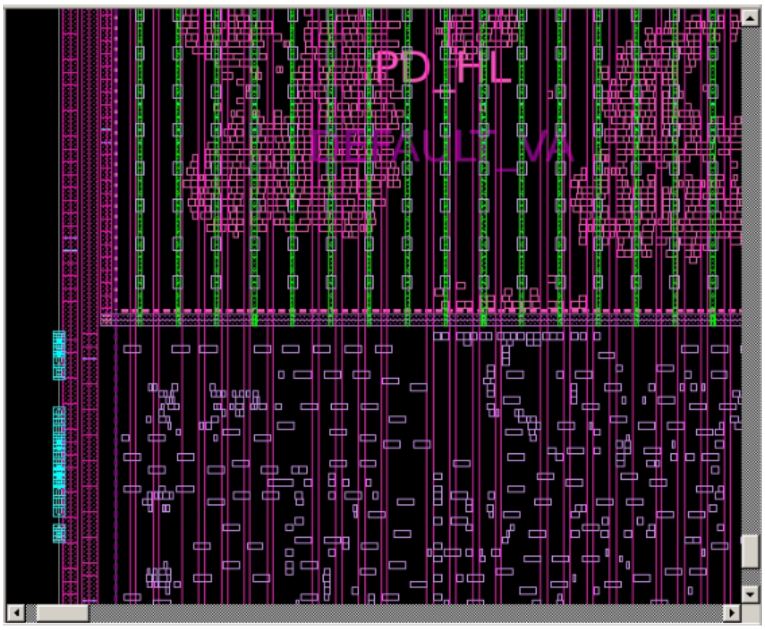


Fig. 4. Layout view of ELM in CTS

5. Conclusion

In conclusion, the power gating was successfully implemented on Extreme Learning Machine (ELM) by following the steps in Very Large-Scale Integration (VLSI). By reporting the power domain and special cells in logic synthesis stage, it showed that the information described in Unified Power Format (UPF) file for power gating was implemented successfully on ELM design. Another verification of implementation of power gating was done using GUI of DC to show the UPF design of the circuit. The functionality and validity of ELM was verified by reporting timing after compilation was done, and the result produced no slack or timing violation, which indicates the design with power gating implementation still perform its function in good condition. After compilation was completed, netlist of the ELM design was generated. Power analysis for power gating was simulated by PrimeTime tool and it decreased 30.43% of power consumption in ELM design. This shows the first objective of this project was achieved, which is decreasing power consumption of ELM design using power gating technique. The power gating was also implemented in the physical design stage of ELM design. Although there are improvements to be made due to the complexity of ELM, the power gating still considered as successfully implemented in physical design stage through a case study of power gating on the simpler design which is square of summation circuit. Hence, the second objective of this project is considered achieved.

References

[1] Martin, D. *Google uses deep learning to design faster, smaller AI chips*. The Register. (2022, March 18).

[2] Huang, Guang-Bin, Hongming Zhou, Xiaojian Ding, and Rui Zhang. "Extreme learning machine for regression and multiclass classification." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42, no. 2 (2011): 513-529. https://doi.org/10.1109/TSMCB.2011.2168604

[3] Peterson, Dustin, and Oliver Bringmann. "Power-Gating Models for Rapid Design Exploration." In *2019 17th IEEE International New Circuits and Systems Conference (NEWCAS)*, pp. 1-4. IEEE, 2019. https://doi.org/10.1109/NEWCAS44328.2019.8961232

[4] Chong Yeam, T. (n.d.). Algorithm Restructuring and Directives Configuration of High-Level Synthesis towards High Performance Extreme Learning Machine Accelerator. *Universiti Teknologi Malaysia.*