**A FLEXIBLE AND ENERGY-EFFICIENT CONVOLUTIONAL NEURAL NETWORK ACCELERATION WITH DEDICATED ISA AND ACCELERATOR**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE NO** |
|  | **ABSTRACT** |  |
|  | **INTRODUCTION** |  |
|  | * 1. General Introduction |  |
|  | * 1. Project Objectives |  |
|  | * 1. Problem Statement |  |
|  | **SYSTEM** **PROPOSAL** |  |
|  | * 1. Existing System |  |
|  | * + 1. Disadvantages |  |
|  | * 1. Proposed System |  |
|  | 2.2.1 Advantages |  |
|  | * 1. Literature Survey |  |
|  | **SYSTEM DIAGRAMS** |  |
|  | * 1. Architecture Diagram |  |
|  | * 1. Flow Diagram |  |
|  | * 1. UML Diagrams |  |
|  | **IMPLEMENTATION** |  |
|  | * 1. Modules |  |
|  | * 1. Modules Description |  |
|  | **SYSTEM** **REQUIREMENTS** |  |
|  | * 1. Hardware Requirements |  |
|  | * 1. Software Requirements |  |
|  | * 1. Software Description |  |
|  | * 1. Testing of Products |  |
|  | **CONCLUSION** |  |
| **7.** | **FUTURE** **ENHANCEMENT** |  |
| **8.** | **SAMPLE** **CODING** |  |
| **9.** | **SAMPLE** **SCREENSHOT** |  |
| **10.** | **REFERENCES** |  |

**ABSTRACT**

Deep learning neural networks have gained much attention in recent research. Excellent results in various domains have proved the usefulness of such algorithms. However, training a deep learning network requires substantial computational effort; therefore, resource-constrained systems like edge devices in the IoT domain still lack full implementations, and training of the network is ofﬂoaded to the cloud. Online or unsupervised training of the network, on the other hand, is often a must if the system has to adjust to possible drift of the environment parameters or there is not enough data available initially. This paper proposes the ﬁrst Xilinx Zynq FPGA (Field Programmable Gate Array) based implementation of the contractive auto encoder (CAE), including training of the network

**INTRODUCTION**

Deep learning (DL) algorithms have been proved to be useful in various domains: image recognition, natural language translation, human activity recognition, and anomaly detection [1], [2], [3]. However, the current state-of-the-art solutions rely on graphical processing units and other general-purpose hardware accelerators. The DL algorithms extract the essential features of the input signal automatically; this enables automatic learning and increases the DL modeling capabilities [4]. Before the deployment, DL algorithms need training, which requires substantial computational power. Therefore, the network is either trained offline, or using the cloud [5]. The broader focus of this work is related to the unsupervised DL algorithms and implementations on resource constrained systems. One class of this kind of methods are auto encoders, which reproduce the input signal to its output. The middle layer of an auto encoder contains compressed features [6], which can be used for different purposes, like data-compression [7]. [8] Describes the framework for FPGA based forward pass execution of various DL networks but does not include the training, which has to be carried out separately. Considering auto encoders, [9] provides the study of an FPGA based sparse stacked auto encoder, but again, it does lack the training. Using high-level synthesis is another approach found in the literature; [10] provides the solution to train stacked auto encoders. However, the proposed solution lacks the training speed and the contraction term. The main contribution of this work is to provide the first hardware-based implementation of the Contractive

**EXISTING SYSTEM**

In existing system, model of autoencoder is one of the most typical deep learning models that have been mainly used in unsupervised feature learning for many applications like recognition, identification and mining. Autoencoder algorithms are compute-intensive tasks. Building large scale autoencoder model can satisfy the analysis requirement of huge volume data. But the training time sometimes becomes unbearable, which naturally leads to investigate some hardware acceleration platforms like FPGA. The software versions of autoencoder often use single-precision or double-precision expressions. But the floating point units are very expensive to implement on FPGA. Fixed-point arithmetic is often used when implementing autoencoder on hardware. But the accuracy loss is often ignored and its implications for accuracy have not been studied in previous works. There are only some works focused on accelerators using some fixed bit-widths on other neural networks models. Our work gives a comprehensive evaluation to demonstrate the fix-point precision implications on the autoencoder, achieving best performance and area efficiency. The method of data format conversion, the matrix blocking methods and the complex functions approximation are the main factors considered according to the situation of hardware implementation. The simulation method of the data conversion, the matrix blocking with different parallelism and a simple PLA approximation method were evaluated in this paper. The results showed that the fixed-point bit-width did have effect on the performance of autoencoder. Multiple factors may have crossed effect. Each factor would have two-sided impacts for discarding the "abundant" information and the "useful" information at the same time. The representation domain must be carefully selected according to the computation parallelism. The result also showed that using fixed-point arithmetic can guarantee the precision of the autoencoder algorithm and get acceptable convergence speed.

* **DISADVANTAGES**
* In the existing system, more inverter gates that are used for memory architecture.
* This may consume more leakage power. And requires relatively high write energy to build up SRAM architecture.

**PROPOSED SYSTEM**

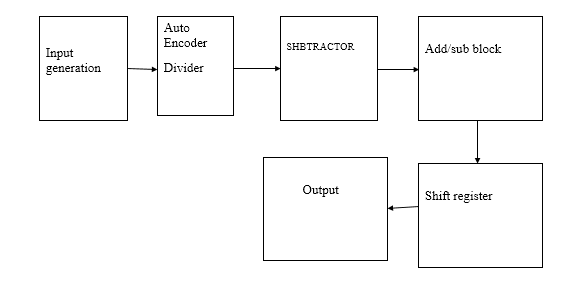
In proposed system Deep learning neural networks have gained much attention in recent research. Excellent results in various domains have proved the usefulness of such algorithms. However, training a deep learning network requires substantial computational effort; therefore, resource-constrained systems like edge devices in the IoT domain still lack full implementations, and training of the network is ofﬂoaded to the cloud. Online or unsupervised training of the network, on the other hand, is often a must if the system has to adjust to possible drift of the environment parameters or there is not enough data available initially. This paper proposes the ﬁrst FPGA (Field Programmable Gate Array) based implementation of the contractive auto encoder (CAE), including training of the network

**ADVANTAGES**

* This technique is to reduce the energy consumption level and to optimize the writing in the SRAM memory functions.
* The proposed system is used to reduce the power consumption level.
* To proposed system is used to reduce the circuit complexity level.

**SYSTEM DESIGNS**

**SYSTEM ARCHITECTURE**



**FLOW DIAGRAM**

GIVEN INPUT DATA BITS

AUTO ENCODER

ADD/SUB BLOCK

WRITE AND READ THE OUTPUT BITS

MEMORY WRITE PROCESS CONTROL BITS

RESULT PERFORMANCE ANALYSIS

**TESTING OF PRODUCT**

System testing is the stage of implementation, which aimed at ensuring that system works accurately and efficiently before the live operation commence. Testing is the process of executing a program with the intent of finding an error. A good test case is one that has a high probability of finding an error. A successful test is one that answers a yet undiscovered error.

Testing is vital to the success of the system.  System testing makes a logical assumption that if all parts of the system are correct, the goal will be successfully achieved.  The candidate system is subject to variety of tests-on-line response, Volume Street, recovery and security and usability test.  A series of tests are performed before the system is ready for the user acceptance testing.  Any engineered product can be tested in one of the following ways.  Knowing the specified function that a product has been designed to from, test can be conducted to demonstrate each function is fully operational.  Knowing the internal working of a product, tests can be conducted to ensure that “al gears mesh”, that is the internal operation of the product performs according to the specification and all internal components have been adequately exercised.

**UNIT TESTING:**

Unit testing is the testing of each module and the integration of the overall system is done.  Unit testing becomes verification efforts on the smallest unit of software design in the module.  This is also known as ‘module testing’.  The modules of the system are tested separately.  This testing is carried out during the programming itself.  In this testing step, each model is found to be working satisfactorily as regard to the expected output from the module.  There are some validation checks for the fields.  For example, the validation check is done for verifying the data given by the user where both format and validity of the data entered is included.  It is very easy to find error and debug the system.

**INTEGRATION TESTING:**

Data can be lost across an interface, one module can have an adverse effect on the other sub function, when combined, may not produce the desired major function.  Integrated testing is systematic testing that can be done with sample data.  The need for the integrated test is to find the overall system performance. There are two types of integration testing. They are:

1. Top-down integration testing.
2. Bottom-up integration testing.

**WHITE BOX TESTING:**

White Box testing is a test case design method that uses the control structure of the procedural design to drive cases.  Using the white box testing methods, we derived test cases that guarantee that all independent paths within a module have been exercised at least once.

**BLACK BOX TESTING:**

* + Black box testing is done to find incorrect or missing function
  + Interface error
  + Errors in external database access
  + Performance errors
  + Initialization and termination errors

In ‘functional testing’, is performed to validate an application conforms to its specifications of correctly performs all its required functions. So this testing is also called ‘black box testing’.  It tests the external behavior of the system.  Here the engineered product can be tested knowing the specified function that a product has been designed to perform, tests can be conducted to demonstrate that each function is fully operational.

**VALIDATION TESTING:**

After the culmination of black box testing, software is completed assembly as a package, interfacing errors have been uncovered and corrected and final series of software validation tests begin validation testing can be defined as many, but a single definition is that validation succeeds when the software functions in a manner that can be reasonably expected by the customer.

# USER ACCEPTANCE TESTING:

User acceptance of the system is the key factor for the success of the system.  The system under consideration is tested for user acceptance by constantly keeping in touch with prospective system at the time of developing changes whenever required.

# OUTPUT TESTING:

After performing the validation testing, the next step is output asking the user about the format required testing of the proposed system, since no system could be useful if it does not produce the required output in the specific format.  The output displayed or generated by the system under consideration.  Here the output format is considered in two ways.  One is screen and the other is printed format.  The output format on the screen is found to be correct as the format was designed in the system phase according to the user needs.  For the hard copy also output comes out as the specified requirements by the user. Hence the output testing does not result in any connection in the system.

**System Implementation:**

Implementation of software refers to the final installation of the package in its real environment, to the satisfaction of the intended users and the operation of the system. The people are not sure that the software is meant to make their job easier.

* The active user must be aware of the benefits of using the system
* Their confidence in the software built up
* Proper guidance is impaired to the user so that he is comfortable in using the application

Before going ahead and viewing the system, the user must know that for viewing the result, the server program should be running in the server. If the server object is not running on the server, the actual processes will not take place.

**User Training:**

To achieve the objectives and benefits expected from the proposed system it is essential for the people who will be involved to be confident of their role in the new system. As system becomes more complex, the need for education and training is more and more important.

Education is complementary to training. It brings life to formal training by explaining the background to the resources for them. Education involves creating the right atmosphere and motivating user staff. Education information can make training more interesting and more understandable.

**Training on the Application Software:**

After providing the necessary basic training on the computer awareness, the users will have to be trained on the new application software. This will give the underlying philosophy of the use of the new system such as the screen flow, screen design, type of help on the screen, type of errors while entering the data, the corresponding validation check at each entry and the ways to correct the data entered. This training may be different across different user groups and across different levels of hierarchy.

**Operational Documentation:**

Once the implementation plan is decided, it is essential that the user of the system is made familiar and comfortable with the environment. A documentation providing the whole operations of the system is being developed. Useful tips and guidance is given inside the application itself to the user. The system is developed user friendly so that the user can work the system from the tips given in the application itself.

**System Maintenance:**

The maintenance phase of the software cycle is the time in which software performs useful work. After a system is successfully implemented, it should be maintained in a proper manner. System maintenance is an important aspect in the software development life cycle. The need for system maintenance is to make adaptable to the changes in the system environment. There may be social, technical and other environmental changes, which affect a system which is being implemented. Software product enhancements may involve providing new functional capabilities, improving user displays and mode of interaction, upgrading the performance characteristics of the system. So only thru proper system maintenance procedures, the system can be adapted to cope up with these changes. Software maintenance is of course, far more than “finding mistakes”.

**Corrective Maintenance:**

The first maintenance activity occurs because it is unreasonable to assume that software testing will uncover all latent errors in a large software system. During the use of any large program, errors will occur and be reported to the developer. The process that includes the diagnosis and correction of one or more errors is called Corrective Maintenance.

**Adaptive Maintenance:**

The second activity that contributes to a definition of maintenance occurs because of the rapid change that is encountered in every aspect of computing. Therefore Adaptive maintenance termed as an activity that modifies software to properly interfere with a changing environment is both necessary and commonplace.

**Perceptive Maintenance:**

The third activity that may be applied to a definition of maintenance occurs when a software package is successful. As the software is used, recommendations for new capabilities, modifications to existing functions, and general enhancement are received from users. To satisfy requests in this category, Perceptive maintenance is performed. This activity accounts for the majority of all efforts expended on software maintenance.

**Preventive Maintenance:**

The fourth maintenance activity occurs when software is changed to improve future maintainability or reliability, or to provide a better basis for future enhancements. Often called preventive maintenance, this activity is characterized by reverse engineering and re-engineering techniques.

**MODULES**

* Input
* CNN Algorithm
* Register
* Data memory block
* Square root
* Accumulator
* Control unit

**LITERATURE SURVEY**

1.Title: Giant Magnetoresistance

Author Name: N. Shirato

The discovery of giant magnetoresistance (GMR) has been a huge impact on our life, especially for mass data storage devices. Initial experiments conducted by Gruberg and Fert are explained. Basic physics of the GMR effect can be explained by the two-current model, which the conduction of a current is consist of two different spin electrons. Details of GMR applications, such as hard-disk read-heads and magnetic memory chips are presented.

In our every day life, it’s inseparable to live without digital data. The discovery of giant magnetoresistance (GMR)in by two french and german scientists, Albert Fert and Peter Grunberg has been dramatically improving our way to live. GMR’s application to the read head of hard discs greatly contributed to the fast rise in the density of stored information and led to the extension of the hard disk technology to consumer’s electronics. For instance, since the introduction of GMR-type sensors as reading elements, in around storage capacities have increased approximately 100 times. Besides in terms of further technological advances, the development of spintronics revealed many other phenomena related to the control and manipulation o f spin currents. Thus basically GMR of the magnetic multilayers opened the way to an efficient control of the motion of the electrons by acting on their spin through the orientation of a magnetization.

The Cr thickness was d0 =1 nm, so that the Fe layers were coupled antiferromagnetically providing an antiparallel alignment of their magnetizations at

zero applied magn etic field. As a reference sample, they also made a single Fe film with thickness d=25 nm in 100 order to measure the anisotropic magnetoresistance (AMR) effect for comparison. Laterally, the samples had the shape of a long strip with contacts at both ends. They have an MR effect both due to the anisotropic effect (negative values) and antiparallel alignment (positive values). Using Mott’s arguments it is straightforward to explain GMR in magnetic multilayers. We con sider the two-current model that a current is consist of each spin-up (blue arrow) and spin-down (red arrow) electrons movement, and a sample is made of combinations of ferromagnetic and non-magnetic metals layers simultaneously, and as-sume that the scattering is strong for electrons with spin an-tiparallel to the magnetization direction, and is weak for elec-trons with spin parallel to the magnetization direction.

On the other hand, the down-spin electrons are scattered strongly within both ferro-magnetic layers, because their spin is antiparallel to the mag-netization of the layers. Since conduction occurs in parallel for the two spin channels, the total resistivity of the multilayer is determined mainly by the highly-conductive up-spin elec-trons and appears to be low

Advantage:

* GMR’s application to the read head of hard discs greatly contributed to the fast rise in the density of stored in-formation and led to the extension of the hard disk technology to consumer’s electronics.

Disadvantage:

* The ferromagnetic layers progressively rotate towards the field, leading to a decrease in the resistance of the multilayer

2.Title: An Overview of Spin-based Integrated Circuits

Author Name: Wang Kang, Weisheng Zhao, Zhaohao Wang, Jacques-Olivier Klein, Yue Zhang, Djaafar

Conventional CMOS integrated circuits suffer from serve power and scalability challenges as technology node scales into ultra-deep-micron technology nodes. Alternative approaches beyond charge-only based circuits. In particular, spin-based devices or integrated circuits show promising merits to overcome these issues by adding the spin freedom of electrons to the electronic circuits. Spintronics has now become a hot topic in both academics and industrials. This paper overviews the status and prospects of spin-based integrated circuits under intense investigation and address particularly their merits and challenges for practical applications.

Many new advanced solutions, e.g., silicon on insulator (SOI), have been proposed recently in order to alleviate this dilemma. They achieve one order of enhancement in some aspects, but they cannot overcome all the problems induced in the deep-micron technology nodes. In addition, the production cost including facility investment becomes very huge. Therefore alternative scaling-independent technologies, i.e., spin-based devices and circuits, for improving the integrated circuit performance have attracted considerable attention to sustain the 0RRUH¶V/DZEH\RQGWKHMOS scaling limit. Many research groups, including academia and industries are undergoing for this purpose.

In the spin-MOSFET devices, the alignment of the drain magnetization is fixed, while that of the source can be changed, so the gate allows current to flow from the source to the drain without modulation. In contrast to the spin-FETs, the cutoff state of the spin-MOSFET is simply achieved by a gate bias condition in the same manner as an ordinary MOS transistor. In both type of devices, the spin is injected from the ferromagnetic source, and then transported through the channel to the drain and electrons with spin aligned with the drain are passed and generate current. Spin transistors provide a potential building element for novel integrated circuits and open a promising path for achieving real all-spin based integrated circuits.

Another severe issue is the poor sense reliability caused mainly by the device mismatch (both MOS and MTJs devices) of the S.As and the intrinsic stochastic switching effects of the MTJs. Different from the memory circuits where complex error correction circuits (ECCs) can be employed, it is difficult to embed them in the logic circuits while keeping fast speed, low area and high power efficiency. Therefore the current efforts that concentrate on this topic are fast-access MTJ development, high-performance S.A design, low-cost and reliable integration process etc.

Advantage:

* They achieve one order of enhancement in some aspects, but they cannot overcome all the problems induced in the deep-micron technology nodes.
* The key challenge for STT-MRAM is to achieve low write current, high speed, good endurance as well as long retention

Disadvantage:

* It is difficult to embed them in the logic circuits while keeping fast speed, low area and high power efficiency.

3.Title: 3D Vertical Dual-Layer Oxide Memristive Devices for Neuromorphic Computing

Author Name: Siddharth Gaba, Patrick Sheridan, Chao Du, and Wei Lu

Dual-layer resistive switching devices with horizontal W electrodes, vertical Pd electrodes and WOx switching layer formed at the sidewall of the horizontal electrodes have been fabricated and characterized. The devices exhibit well-characterized analog switching characteristics and small mismatch in electrical characteristics for devices formed at the two layers. The three-dimensional (3D) vertical device structure allows higher storage density and larger connectivity for neuromorphic computing applications. We show the vertical devices exhibit potentiation and depression characteristics similar to planar devices, and can be programmed independently with no crosstalk between the layers

The vertical 3D RRAM structure or sidewall type of structure has several advantages over the traditional crosspoint-type of device structure. In a conventional cross point device, the active area dimensions are completely defined by lithography. As devices scale, the cost of lithography steps increases drastically. However, in sidewall devices at least one active device dimension is not critically dependent on lithography. The top (vertical) electrode is still defined by lithography in both conventional crosspoint devices and in sidewall devices. In the latter case, however, the active area dimension is determined by the thickness of a deposited film which can be precisely controlled to the atomic level, as opposed to the lithographically defined dimension of the bottom electrode. Additionally, since deposition thicknesses can be controlled to a much better extent than defined by lithography, device to device variation can be improved.

The tungsten and silicon dioxide depositions were then repeated to form the dual-layer horizontal electrode stack (inset). Next, photolithography and reactive ion etching (RIE) were used to pattern the film stack. To form the tungsten oxide (WOX) switching layer, the sample was annealed in an oxygen rich ambient at 375°C at atmospheric pressure for 60 seconds in a JetFirst 150 RTP system. CMOS compatible, dual-layer vertical tungsten oxide resistive switching devices were demonstrated. The devices show well-defined incremental resistance switching behavior and good endurance exceeding 10,000 potentiation/depression cycles. The devices can be programmed with less than 10% mismatch and no apparent crosstalk. This scalable architecture is well suited for development of analog memory and neuromorphic systems. The conductance change ratio may be further increased by optimizing the stack etch, post etch cleans and the oxidation conditions and is the subject of further studies

Advantage:

* Since deposition thicknesses can be controlled to a much better extent than defined by lithography, device to device variation can be improved.

Disadvantage:

* The total device resistance is dominated by the oxygen-vacancy poor region near the horizontal electrode

4.Title: AN EFFICIENT H-TREE BASED CLOCK TREE DESIGN USING AGGLOMERATIVE CLUSTERING ALGORITHM

Author Name: Radhika V Murugasami R

Power consumption has become an important issue in high-performance circuits. Several techniques are used to reduce total power of a chip, such as multiple supply voltages, clock gating, and clock-tree minimization. Minimizing the size of a clock tree is known as an effective approach to reduce the power dissipation in modern circuit designs. Clock tree has different types of sinks like flip flops and pulsed latches. In this system latches are connected to the buffer at a limited position, each latch can communicate to nearby buffer this form the forms the tree based network. The clock tree with mixture of sinks is constructed to reduce power dissipation, but the load level of clock tree is increased. In the proposed system, load level of clock tree, the power and clock skew are reduced by using agglomerative clustering algorithm in the H-tree based architecture. Multi-corner multi-mode (MCMM) Clock tree synthesis (CTS) is used in H-Tree based topology to reduce the load level in clock tree. Experimentally load level of clock tree, the power and clock skew of the system is improved by using the agglomerative clustering algorithm in H tree based architecture. Xilinx tool will be used to carry out the proposed system. The proposed system will be implemented using Hardware Description Language (VHDL). Keywords: Clock tree design, dynamic power reduction, Flip flops, Agglomerative clustering algorithm.

A clustering algorithm that uses a minimum spanning tree to estimate the interconnect Capacitance and reduce the total wire capacitance in the routing Stage. However, existing methods of clock-tree minimization are primarily based on flip-flops and focus on wire length Minimization alone, which may limit achievable power Savings. In current circuit designs, the most common storage element is a D-type flip-flop that consists of two latches (master and Slave) triggered by a clock signal. This type of design makes it easier to apply static timing analysis (STA) for timing Verification. As transistor counts of a flip-flop are two times than that of a single latch, latches are superior to flip-flops in terms of area, transition time, and power dissipation. However, it is difficult to perform STA on latch-based circuits because of data transparency. A pulsed-latch-based design style was adopted for dynamic power reduction. Pulsed latches are latches triggered by a brief clock signal Generated from a pulse generator. When the pulse clock Waveform triggers a latch, the latch is synchronized with the Clock and its timing behavior is similar to an edge-triggered Flip-flop

To prevent pulse degradation, the tolerable load of a pulse generator and the number of pulsed latches driven by a pulse generator were considered during pulse-generator insertion. To further reduce the power dissipation of pulse generators, we enabled multi-type pulse generators and identified the pulse generators with suitable size to drive pulsed latches.

Advantage:

* Minimizing the size of a clock tree is known as an effective approach to reduce the power dissipation in modern circuit designs

Disadvantage:

* It is difficult to perform STA on latch-based circuits because of data transparency

5.Title: Designing and Analysis of 8 Bit SRAM Cell with Low Subthreshold Leakage Power

Author Name: Atluri.Jhansi rani, K.Harikishore, Fazal Noor Basha,V.G.Santhi Swaroop

The power consumption is major concern in Very Large Scale Integration (VLSI) circuit design and reduce the power dissipation is challenging job for low power designers. . International technology roadmap for semiconductors (ITRS) reports that “leakage power dissipation” may come to dominate total power consumption. The sub-threshold leakage power is the main reason to increase the leakage power. So there is some techniques to reduce this leakage power like sleep approach, stack & some new techniques like, sleepy–stack, leakage feedback approach and sleepy keeper techniques which reduces leakage current while saving exact logic state. As the technology increases integration density of transistors increases, power consumption has become a major concern in today’s processors and SoC designs. Considerable attention has been paid to the design of low power and high-performance SRAMs as they are critical components in both handheld devices and high performance processors. In this paper we design 8 bit S-RAM by using the leakage current reduction techniques. The proposed circuits were designed in 0.18µm CMOS/VLSI technology with-in Micro-Wind tool, and measure power consumption for design approaches, and we achieves up to nearly 50% less power consumption than existing basic SRAM.

The art of power analysis and optimization of integrated circuits used to be a specialty in analog circuit design. Power dissipation of VLSI chips is traditionally a neglected subject. In the past the device density and operating frequency were low enough that it was a constraining factor in the chips. As the technology varies, more transistors, faster and smaller than their predecessors, which leads to the growth in operating frequency and processing per capacity leads to increase in power consumption. There are two types of power dissipation in CMOS Circuits: Dynamic and Static. Dynamic power is caused by switching activities of the circuit and most significant source of dynamic power dissipation in CMOS circuits is the charging and discharging of the capacitance. Static Power dissipation is related to the logical states of the circuits rather than switching activities.

In CMOS logic, leakage current is the only source of static power dissipation. Currently, sub-threshold leakage seems to be the dominant contributor to overall leakage power. Another possible contributor to leakage power is gate-oxide leakage. A possible solution widely reported is the potential use of high k (high dielectric constant) gate insulators. In any case, this papers targets reduction of the sub-threshold leakage component of static power consumption; other approaches should be considered for reduction of gate oxide leakage. Do please note, however, that all results reported in this paper include all sources of leakage power. With application of dual threshold voltage (Vth) techniques, the sleep, zigzag and sleepy stack approaches result in orders of magnitude sub-threshold leakage power reduction.

Advantage:

* Each technique provides an efficient way to reduce leakage power,

Disadvantage:

* SRAM Architecture is used for low power designs and these designed techniques are used for high performance and low power applications

**Title 6: Low-power Hybrid CAM for High speed route Lookup Engines**

**Author**: mr. K. Suresh kumar, dr.y. Rajasree rao, dr. K.manjunathachari

**Year: 2013**

Content-addressable memory (CAM) is a hardware table that can compare the search data with all the stored data in parallel. Due to the parallel comparison feature where a large amount of transistors are active on each lookup, however, the power consumption of CAM is usually considerable. This paper presents a hybrid-type CAM design which aims to combine the performance advantage of the NOR-type CAM with the power efficiency of the NAND-type CAM. In our design, a CAM word is divided into two segments, and then all the CAM cells are decoupled from the match line. The experimental results show that the hybrid-type CAM can reduce the search energy consumption by roughly 89% compared to the traditional NOR-type CAM. Because the hybrid-type CAM provides a fast pulldown path to speed up the lightweight match line discharge, the search performance of our design is even better than that of the traditional NOR-type CAM.

A Content Addressable Memory (CAM) compares input search data against a table of stored data, and returns the address of the matching data. CAMs have a single clock cycle throughput making them faster than other hardware- and software-based search systems. CAMs can be used in a wide variety of applications requiring high search speeds. These applications include parametric  
curve extraction, Hough transformation, Huffman coding/decoding, Lempel–Ziv compression, and image coding. The primary commercial application of CAMs today is to classify and forward Internet protocol (IP) packets in network routers.

In networks like the Internet, a message such an as e-mail or a Web page is  
transferred by first breaking up the message into small data packets of a few hundred bytes, and, then, sending each data packet individually through the network. These packets are routed from the source, through intermediate nodes of the network (called routers), and reassembled at the destination to reproduce the  
original message. The function of a router is to compare the destination address of a packet to all possible routes, in order to choose the appropriate one. A CAM is a good choice for implementing this lookup operation due to its fast search capability.

However, the speed of a CAM comes at the cost of increased silicon area and power consumption, two design parameters that designers strive to reduce. As CAM applications grow, demanding larger CAM sizes, the power problem is further exacerbated. Reducing power consumption, without sacrificing speed or area, is the main thread of recent research in large-capacity CAMs. In this paper, we survey developments in the CAM area at two levels: circuits and architectures. Before providing an outline of this paper at the end of this section, we first briefly introduce the operation of CAM and also describe the CAM application of packet forwarding.

The number of bits in a CAM word is usually large, with existing implementations ranging from 36 to 144 bits. A typical CAM employs a table size ranging between a few hundred entries to 32K entries, corresponding to an address  
space ranging from 7 bits to 15 bits. Each stored word has a matchline that indicates whether the search word and stored word are identical (the match case) or are different (a mismatch case, or miss). The matchlines are fed to an encoder that generates a binary match location corresponding to the matchline that is in the match state. An encoder is used in systems where only a single match is expected.

In CAM applications where more than one word may match, a priority encoder is used instead of a simple encoder. A priority encoder selects the highest priority matching location to map to the match result, with words in lower address locations receiving higher priority. In addition, there is often a hit signal (not  
shown in the figure) that flags the case in which there is no matching location in the CAM. The overall function of a CAM is to take a search word and return the matching memory location. One can think of this operation as a fully programmable arbitrary mapping of the large space of the input search word to the smaller space of the output match location.

A CAM search operation begins with precharging all matchlines high, putting them all temporarily in the match state. Next, the search line drivers broadcast the search data, 01101 in the figure, onto the search lines. Then each CAM core cell compares its stored bit against the bit on its corresponding search lines. Cells with matching data do not affect the matchline but cells with a   
mismatch pull down the matchline. Cells storing an X operate as if a match has occurred. The aggregate result is that matchlines are pulled down for any word that has at least one mismatch. All other matchlines remain activated (precharged high). the two middle matchlines remain activated, indicating a match, while the other matchlines discharge to ground, indicating a mismatch..

In this paper, we have developed a hybrid-type CAM design, in which we decouple all the CAM cells from the match line, and provide a fast path to accelerate the search operation. With a marginal area overhead, our design not only largely reduces the search power consumption but also improves the search  
performance.

**Advantages**: it not only largely reduces the search power, but also improves the match delay.

**Disadvantages:** It is not a feasible solution because of its long match delay.

**Title 7: design of low power pre-computation based cam using xor and gate block selection scheme**

**Author:** Pallavi Shivatare, V.G.Raut

**Year: 2015**

Content-addressable memory (CAM) is a special type of computer memory which provides a fast data search operation in single clock cycle. Content-addressable memory (CAM) is frequently used in applications such as lookup tables, databases, associative computing, and networking. However the high speed of CAM increases the power consumption. This paper presents low power techniques i.e. One’s count, XOR approach and Gate block selection to improve power efficiency of pre-computation based CAM. In this experiment VHDL modeling is used and implemented using Xilinx to estimate the power consumption. Compared with the ones count PB-CAM system the experimental result shows that proposed approach can achieve 30 % power reduction.

Content addressable memory (CAM) is the special type of memory in which data can be identified by its content instead of its address. CAM compares input search data with stored data, and returns the address of the matching data. CAM has fast search capability. CAMs can be used in a wide variety of applications such as asynchronous transfer mode (ATM), communication networks, databases, lookup tables, data compression. Due to its vast number of comparison operation CAM consumes large amount of power.

Content addressable memory (CAM) provides a fast data search function by comparing the search data with all of the stored data in a single clock cycle. CAMs can be used in a wide variety of applications requiring high search speeds. However due to high speed of a CAM, silicon area and power consumption increases. For lager CAM the power problem further increases. To reduce the power consumption the CAM can be modified at architecture level.

We can compare CAM to RAM. RAM produces the data for a given address. But in case of CAM, it produces an address for a given data word. For RAM, data search is done serially. Thus, finding a particular data word can take many cycles. In case of CAM searches all addresses in parallel and produces the address storing a particular word. CAM supports writing "don't care" bits into words of the memory. The don't care bit can be used as a mask for CAM comparisons. The output of the CAM can be encoded. It ensures duplicate data is not written into the CAM.

Each stored word has matchline which indicates that whether the search data and stored data are same or different. Then matchlines are connected to encoder. The encoder gives the match location corresponding to match line which in match state. CAM does this searching operation at very high speed. So it requires large  
amount power. In order to reduce the power consumption we have introduced PB-CAM (pre computation-Based CAM) architecture.

The disadvantage of one‟s-count technique is that from parameter 5 to 9 the number of data related to the same parameter is around 2000-3000. Hence comparison operations also increased simultaneously which will also increase the power consumption. Ones-count PB-CAMs fail to reduce the number of comparison operations in the second part. As it can be seen in Table I, random input patterns for the ones-count approach demonstrate the Gaussian distribution characteristic. Here the Gaussian distribution will limit any further reduction of the  
comparison operations in PB-CAMs. So, we replaced Block-XOR technique for reducing the amount power consumption.

In XOR gate PB-CAM parameter extractor we consider only XOR logic gate. So to make the parameter extractor more useful for specific data types, we consider different characteristic of logic gates. The proposed parameter extractor architecture. There are several partition block and each block contains several subblocks G0-G6. Each sub-block stands for different logic gates. Output bit is computed using synthesized logic operation for each of these sub-blocks. Then the output bits will become the parameter for data comparison process. The main objective is to select the proper logic gates so that the parameter can reduce the  
number of data comparison operations.

In this paper low power pre-computation based content addressable memory (PB-CAM) is simulated in VHDL. Simulation results and mathematical analysis shows that the XOR and gate-block selection techniques can effectively save power as compared to one‟s count. It also reduces the number of comparison operations. From power report it shows that power consumption is also reduced. This PB-CAM performs data searching operation in exactly one clock cycle. So it is flexible and adaptive for the low power and high speed search applications such as asynchronous transfer mode (ATM), communication networks, databases, lookup tables, data compression. The power can be further reduced by using XNOR approach instead of XOR approach for parameter extractor design of CAM.

**Advantages**: It can be used as a mask for CAM comparisons

**Disadvantages:** .From parameter 5 to 9 the number of data related to the same  
parameter is around 2000-3000.

**Title 8: low power concept for content addressable memory (cam) chip  
design**

**Author:** Dejan Georgiev

**Year:** 2013

A Content Addressable Memory (CAM) is a memory unit that performs single clock cycle content matching instead of addresses. CAM's are vast used in look-up table functions , network routers and cache controllers. Since basic lookups are performed over all the stored memory information there is a high power dissipation. In reality there is always trade-offs between power consumption, area used and the speed. Here is presented an conceptual abstraction for content addressable memory chip at architecture level with reduced power requirements based on combination and modifications of power saving techniques.

Content Addressable Memories (CAM) are fast data parallel search circuits. Unlike standard memory circuits , for example Random Access Memory (RAM) data search is performed against all the stored information in single clock cycle. In fact CAM is outgrowth of RAM. While CAM's are widely used in many applications like memory mapping, cache controllers for central processing unit, data compression and coding etc. it primary application is fast Internet Protocol (IP) package classification and forwarding at high speed network routers and processors. IP routing is accomplished by examination of the protocol header fields i.e. the originating and destination address, the incoming and outgoing ports etc. against stored information in the routing tables. If a match is registered the package is forwarded towards the port(s) defined in the table. On very high speed networks and huge traffic volume the task is to be performed in fast and massive parallelism. However, managing high speeds and large lookup tables requires silicon  
area and power consumption. The power dissipation , silicon area and the speed are three major challenges for designers. Since there is always trade-off between them , reducing one without sacrificing the others is a is the main treat in recent research for large CAMs

A basic CAM cell function could be observed as twofold: bit storage as in RAM and bit comparison which is unique to CAM. At transistor i.e. circuit level CAM structure implemented as NAND-type or NOR-type and its variants has  
been explained. But at architectural level bit storage uses simple (S)RAM cell and comparison function is equivalent to XOR i.e. XNOR logic operation. Thus our elementary chip cell design is abstracted as a cross product of SRAM and XNOR circuits. Figure 1 represents the logical symbol and the circuit compilation.

The input signal is one bit value from the search data register i.e. the input word to be compared against all the values stored in CAM arrays or the value to be stored in the CAM cell. Cell enable signal allows or prevent comparison i.e.  
matching process meaning XOR-ing the stored bit value in the Flip-flop and the input bit. It should be mentioned the extended truth table of three-state buffer presented on Table 1 where x represents the input signal and y is the output  
signal. "Z" denotes high-impedance or practically disconnected line or switch on/off.

It not only requires huge memory area but unnecessary power lost in  
bit comparisons even if pre-charge or pipeline method is used. For example, for fixed source IP and destination IP addresses computation will be performed over all the first 64 bits regardless of residual bit fields. In that regards an  
improvement can be achieved by statistical pre-computation along word's bits.  
Pre-computation stores some extra bits derived from the stored word and it is used in the initial search before the search of main word.

Selective pre-charge scheme basically divides the mach line in two segments. In general following the same concept it can be divided in many number of segments thus forming a pipeline. If any stage is miss the subsequent stages are shut off resulting in power saving. The drawback of this scheme are the increased latency and area overhead due to the pipeline stages. Here is shown a power saving design scarifying the speed i.e. increased delay but retaining the same circuit area. The basic idea behind the concept is the segmentation in the mach line in a manner that every CAM cell form a segment for its own as it is presented.

The main benefit for the proposed scheme comes from implementation with CAM cells. Namely, the output of a cell is simply the cell enable signal for the successive bit comparison thus mitigating extra gates to transfer the results from the cells. The disadvantage is increased propagation delay that comes from the three-state buffer and XNOR gate at each cell. Typical CAM consist words length ranging from 36 to 144 bits and in practice it should be acceptable delay value. It should be noted that one cell segmentation approach presented here is a conceptual  
view rather than real power saving scheme that can be achieved on circular level.

**Advantages:** It is used in look-up table functions , network routers and cache controllers.

**Disadvantages:** increased propagation delay that comes from the three-state  
buffer and XNOR gate at each cell.

**Title 9: The Efficient Architecture Methods for low power Content Addressable Memory- Survey**

**Author:** SUBHA.M

**Year: 2010**

Content addressable memory (CAM) or associative memory, is a storage device, which can be addressed by its own contents. This paper presents the survey of CAM low power techniques at architecture level. The proposed method of CAM at architecture level consumes less power than all other previous method. This architecture is designed using Tspice-0.18um technology.

Most memory devices store and retrieve data by addressing specific memory locations. As a result, this path often becomes the limiting factor for systems that  
rely on fast memory accesses. The time required to find an item stored in memory can be reduced considerably if the item can be identified for access by its content rather than by its address. A memory that is accessed in this way is called content-addressable memory (CAM). To achieve an effective function of data searching, the data comparison architecture of CAMs is usually implemented in parallel operation structure. However, due to parallel process characteristic, power consumption is always an important Concern when designing CAM circuitry. (i.e.)Content addressable memories simultaneously compare an input word to all the contents of memory and return the address of matching locations. CAMs  
with large capacities speed up the operation of search intensive tasks such as packet Forwarding and classification in routers, database lookups, and compression.

The main challenge in CAM design is to reduce power while maintaining speed and low area. Therefore many articles have been devoted to the Study of CAM for low-power, in which power reduction has focused on the circuit and architecture domains. depicts the basic CAM circuit structure of a CAM word which is made of CAM cells. A CAM cell compares its stored bit against its  
corresponding search bit provided on the search-line (SL). The combined search result for the entire word is generated on the match-line (ML). The match-line  
Sense amplifier (MLSA) senses the state of the ML and outputs a full logic swing signal.

To minimize power consumed during the comparison, one of the best approaches is to reduce most of the comparison operations. Based on this idea, a novel architecture is developed for low-power CAM circuit Design called PB-CAM. In precomputation technique, the output logic values of the circuit are pre computed one clock cycle before they are required. The precomputed logic values are used in the following clock cycle to reduce the switching at the internal node of a circuit.

In the data writing operation, the parameter extractor extracts the parameter of the input data, and then stores the input data and its parameter into the data memory and the parameter memory, respectively. In the data searching operation, in order to reduce the large amount of comparison operations, the operation is separated into two comparison processes. In the first comparison process, the  
parameter extractor extracts the parameter of the input data, and the parameter comparison circuits then compare the parameter of the input data with all  
parameters stored in the parameter memory in parallel.

To minimize power consumed during the comparison, one of the best approaches is to reduce most of the comparison operations. Based on this idea, a novel architecture is developed for low-power CAM circuit Design called PB-CAM. In precomputation technique, the output logic values of the circuit are pre computed one clock cycle before they are required. The precomputed logic values are used in the following clock cycle to reduce the switching at the internal node of a circuit.  
 The data related to this stored parameter concurrently mismatches the input data, if the stored parameter mismatches the parameter of the input data. Otherwise, the data related to this stored parameter has yet to be identified. Using the first comparison process results, the input data is only compared with those unidentified data to identify any match in the second comparison process.

Therefore, the design idea of the parameter extractor is to filter as many unmatched data in the parameter comparison process with the probable shortest bit length of the parameter. Some functions are used to realize the parameter extraction in the proposed CAM architecture, such as ones count function, parity function, and remainder function. In next section, the PB-CAM architecture adopts ones count function to perform the parameter extraction, because the ones count function not only filters a large amount of unmatched data with a small bit length, but also reduces the transistor count of the proposed PBCAM cell to seven transistors.  
 Based on the two comparison processes, if majority parts of the stored parameter mismatch the parameter of the input data, then the number of comparisons in the second comparison process are largely reduced. The function of this parameter comparison process is just like filtering; it filters majority parts of unmatched data in the first comparison process and then reduces most of the comparisons in the second comparison process.

**Advantages:** It can reduce the number of comparison operations

**Disadvantages:** it using full adder circuit .so it consumes huge power and hardware consumption.

**Title 10: Low Power Design of Pre Computation-Based Content-Addressable Memory**

**Author:** SK.Khamuruddeen, S.V.Devika, V Rajath, Vidhan Vikram Varma

**Year: 2014**

Content-addressable memory (CAM) is a special type of computer Memory used in certain very high speed searching applications. It is also known as associative memory, associative storage, or associative array. Content-addressable memory (CAM) is frequently used in applications, such as lookup tables, atabases, associative computing, and networking, that require high-speed searches due to its ability to improve application performance by using parallel comparison to reduce search time. Although the use of parallel comparison results in reduced search time, it also significantly increases power consumption. In this paper, we propose a  
Block-XOR approach to improve the efficiency of low power pre computation- based CAM (PB-CAM). Compared with the onescount PB-CAM system, the experimental results show that our proposed approach can achieve on average 30% in power reduction and 32% in power performance reduction. The major contribution of this paper is that it presents practical proofs to verify that our  
proposed Block-XOR PB-CAM system can achieve greater power reduction without the need for a special CAM cell design. This implies that our approach is more flexible and adaptive for general designs.

However, the speed of a CAM comes at the cost of increased silicon area and power consumption, two design parameters that designers strive to reduce. As CAM applications grow, demanding larger CAM sizes, the power problem is further exacerbated. Reducing power consumption, without sacrificing speed or area, is the main thread of recent research in large-capacity CAMs.

Development in the cam area is surveyed at two levels: circuits and architectures levels. We can compare CAM to the inverse of RAM. When read, RAM produces the data for a given address. Conversely, CAM produces an address for a given data word. When searching for data within a RAM block, the search is performed serially. Thus, finding a particular data word can take many cycles. CAM searches all addresses in parallel and produces the address storing a particular word. CAM supports writing "don't care" bits into words of the memory. The don't care bit can be used as a mask for CAM comparisons; any bit set to don't care has no effect on matches.

In addition, there is often a hit signal (not shown in the figure) that flags the case in which there is no matching location in the CAM. The overall function of a CAM is to take a search word and return the matching memory location. One can think of this operation as a fully programmable arbitrary mapping of the large space of the input search word to the smaller space of the output match location. The operation of a CAM is like that of the tag portion of a fully associative cache. The tag portion of a cache compares its input, which is an address, to all addresses stored in the tag memory. In the case of match, a single match line goes high,  
indicating the location of a match. Many circuits are common to both CAMs and caches; however, we focus on large capacity CAM s rather than on fully associative caches, which target smaller capacity and higher speed.

Next, the search line drivers broadcast the search word onto the differential search lines, and each CAM core cell compares its stored bit against the bit on its corresponding search lines. Match lines on which all bits match remain in the pre charged-high state. Matchlines that have at least one bit that misses, discharge to ground. The MLSA then detects whether its match line has a matching condition or miss condition. Finally, the encoder maps the match line of the matching location to its encoded address.

Since content addressable memory (CAM) is frequently used in applications, that require high-speed searches, and because of its ability to improve application performance by using parallel comparison, it results in reduced search time. But it also significantly increases power consumption. So the main CAM-design challenge is to reduce power consumption associated with the large amount of parallel active circuitry, without sacrificing speed or memory density.

The PB-CAM exploits this characteristic to reduce the comparison operations, thereby saving power. Therefore, the parameter extractor of the PB-CAM is critical, because it determines the number of comparison operations in the second part. So, the parameter extractor plays a significant role since this circuit determines the number of comparison operations required in the second part. Therefore, the design goal of the parameter extractor is to filter out as many unmatched data as possible to minimize the required number of comparison operations in the second part. Two parameter extractors are discussed, namely One‟s count parameter extractor and Block-XOR parameter extractor.

**Advantages**: It is used in certain very high speed searching applications.

**Disadvantages:** it also significantly increases power consumption.

**FEASIBILITY STUDY**

The feasibility study is carried out to test whether the proposed system is worth being implemented. The proposed system will be selected if it is best enough in meeting the performance requirements.

The feasibility carried out mainly in three sections namely.

**•** Economic Feasibility

• Technical Feasibility

• Behavioral Feasibility

**Economic Feasibility**

Economic analysis is the most frequently used method for evaluating effectiveness of the proposed system. More commonly known as cost benefit analysis. This procedure determines the benefits and saving that are expected from the system of the proposed system. The hardware in system department if sufficient for system development.

**Technical Feasibility**

This study center around the system’s department hardware, software and to what extend it can support the proposed system department is having the required hardware and software there is no question of increasing the cost of implementing the proposed system. The criteria, the proposed system is technically feasible and the proposed system can be developed with the existing facility.

**Behavioral Feasibility**

People are inherently resistant to change and need sufficient amount of training, which would result in lot of expenditure for the organization. The proposed system can generate reports with day-to-day information immediately at the user’s request, instead of getting a report, which doesn’t contain much detail.

**System Implementation**

Implementation of software refers to the final installation of the package in its real environment, to the satisfaction of the intended users and the operation of the system. The people are not sure that the software is meant to make their job easier.

* The active user must be aware of the benefits of using the system
* Their confidence in the software built up
* Proper guidance is impaired to the user so that he is comfortable in using the application

Before going ahead and viewing the system, the user must know that for viewing the result, the server program should be running in the server. If the server object is not running on the server, the actual processes will not take place.

User Training

To achieve the objectives and benefits expected from the proposed system it is essential for the people who will be involved to be confident of their role in the new system. As system becomes more complex, the need for education and training is more and more important. Education is complementary to training. It brings life to formal training by explaining the background to the resources for them. Education involves creating the right atmosphere and motivating user staff. Education information can make training more interesting and more understandable.

Training on the Application Software

After providing the necessary basic training on the computer awareness, the users will have to be trained on the new application software. This will give the underlying philosophy of the use of the new system such as the screen flow, screen design, type of help on the screen, type of errors while entering the data, the corresponding validation check at each entry and the ways to correct the data entered. This training may be different across different user groups and across different levels of hierarchy.

Operational Documentation

Once the implementation plan is decided, it is essential that the user of the system is made familiar and comfortable with the environment. A documentation providing the whole operations of the system is being developed. Useful tips and guidance is given inside the application itself to the user. The system is developed user friendly so that the user can work the system from the tips given in the application itself.

System Maintenance

The maintenance phase of the software cycle is the time in which software performs useful work. After a system is successfully implemented, it should be maintained in a proper manner. System maintenance is an important aspect in the software development life cycle. The need for system maintenance is to make adaptable to the changes in the system environment. There may be social, technical and other environmental changes, which affect a system which is being implemented. Software product enhancements may involve providing new functional capabilities, improving user displays and mode of interaction, upgrading the performance characteristics of the system. So only thru proper system maintenance procedures, the system can be adapted to cope up with these changes. Software maintenance is of course, far more than “finding mistakes”.

Corrective Maintenance

The first maintenance activity occurs because it is unreasonable to assume that software testing will uncover all latent errors in a large software system. During the use of any large program, errors will occur and be reported to the developer. The process that includes the diagnosis and correction of one or more errors is called Corrective Maintenance.

Adaptive Maintenance

The second activity that contributes to a definition of maintenance occurs because of the rapid change that is encountered in every aspect of computing. Therefore Adaptive maintenance termed as an activity that modifies software to properly interfere with a changing environment is both necessary and commonplace.

Perceptive Maintenance

The third activity that may be applied to a definition of maintenance occurs when a software package is successful. As the software is used, recommendations for new capabilities, modifications to existing functions, and general enhancement are received from users. To satisfy requests in this category, Perceptive maintenance is performed. This activity accounts for the majority of all efforts expended on software maintenance.

Preventive Maintenance

The fourth maintenance activity occurs when software is changed to improve future maintainability or reliability, or to provide a better basis for future enhancements. Often called preventive maintenance, this activity is characterized by reverse engineering and re-engineering techniques

**SYSTEM REQUIREMENTS**

**Hardware Requirement:**

* Pentium IV – 2.7 GHz
* 1GB DDR RAM
* 250Gb Hard Disk

**Software Requirement:**

* Operating System : Windows XP
* Tool : Vivado ISE 14.7

COMPARISON TABLE:

|  |  |  |
| --- | --- | --- |
|  | EXISTING SYSTEM | PROPOSED SYSTEM |
| TOTAL POWER (µW) | 137.4 | 95 |
| TRANSISTOR COUNT (µm^2) | 1000 | 150 |
| DELAY TIME | 1.2ns | 991.71ps |

**CONCLUSION**

* The proposed architecture uses node-level parallelism. For back-propagation, additional parallelism was achieved by maximally reusing the computational results. The functionality of the solution was verified using the downscaled MNIST dataset The 38μs total execution time for a forward pass and training yields to a maximum of 26KS input rate.

**REFERENCES**

[I] Y. Taur and 1. H. Ning, "Fundamentals of Modern VLSi Devices", New York, USA: Cambridge University Press, 1998, ch. 3, pp. 120-128.

[2] Kim. N, Austin. T, Baauw.D, Mudge. T, Flautner. K, HU. J, Irwin. M, Kandemir.M, and Narayanan.V, "Leakage Current: Moore's Law Meets Static Power", iEEE Computer, vol. 36, pp. 68-75, December 2003.

[3] International Technology Roadmap for Semiconductors by Semiconductor Industry Association, 2002. [Online] Available: http://public.itrs.net.

[4] Soumya Gadag, Raviraj D. Chougla, "Design and Analysis of 6T SRAM Cell with Low Power Dissipation", International Journal of Engineering Research and Application (IJERA), vol. 2, Issue 6, pp. 1695-1698, Nov.-Dec. 2012.

[5] Tajrian Izma, Parag Barua, Md. Rejaur Rahman, and Priyanka Sengupta, "Novel Approaches to Low Leakage and Area Efficient VLSI Design" Ph.D thesis, Dept. Electrical and Electronics Eng., BARC Univ., Dhaka, August 2011.

[6] Neeraj Kr. Shukla, Shilpi Birla, R.K Singh, and Manisha Pattanaik, "Speed and Leakage Power Trad-off in Various SRAM Circuits", international Journal of Computer and Electrical Engineering (IJCEE), Singapore, vol. 3, No.2, Apr. 20ll, pp. 244-249.

[7] Keivan Navi, Roshanak Zabihi, Majid Haghparast, Touraj Nikobin, "A Noval Mixed Current and Dynamic Voltage Full Adder", World Applied Sciences Journal, vol. 4, no.2, pp. 289-294,2008.

[8] A Deepak Lourts, and L. Dhulipalla, "Design and implementation of 32nm FINFET based 4\*4 SRAM cell array using I-bit 6T SRAM", International Conference on Nanoscience, Engineering and Technology (ICONSET), pp. 177-180,28-30 November 2011.

[9] Kavita Khare, Nilay Khare, Vijendra Kumar Kulhade, and Pallavi Deshpande, "VLSI Design and Analysis of Low Power 6T SRAM Cell Using Cadence Tool", International Conference on Semiconductor Electronics (ICSE), pp. 117-121, Malaysia, November 2008.

[10] S. Narendra, V. DE, S. Borkar, D.A Antoniadis, and AP. Chandrakasan, "Full-Chip Subthreshold Leakage Power Prediction and Reduction Techniques for Sub-0.18um CMOS," iEEE Journal of Solid-State Circuits, vol. 39, no. 2, pp. 501-510, February 2004.

[II] Z. Chen, M. Johnson, L. Wei and K.Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stack", international Symoposium on Low Power Electronics and Design, pp. 239-244, August 1998.

[12] S. Mutoh, TDouseki, and Y.Matsuya, "I V Power Supply High-speed Digital Circuit Technology with Multithreshold-Voltage CMOS," iEEE J. Solid-state Circuits, vol. 30, pp. 847-854, Aug., 1995.

[13] Atluri. Jhansi rani, K. Harikishore, Fazal Noor Basha, and V.G. Santhi Swaroop, "Designing and Analysis of 8 Bit SRAM Cell with Low Subthreshold Leakage Power" International Journal of Modern Engineering Research (IJMER), vol.2, Issue.3, pp. 733-741, May­June 2012.

[14] M.Powell, S.H. Yang, B. Falsafi, K. Roy and TN. Vijay Kumar, "Gated-Vdd: A Circuit Technique to Reduce Leakage in Deep Submicron Cache Memories", International Symposium on Low Power Electronics and DeSign, pp. 90-95, July 2000.

[15] Andrei Povlov, O. Semenov and Manoj Sachdev, "Sub-quarter micron SRAM cells stability in low-voltage operation: a comparative analysis", iEEE International Integrated Reliability Workshop Final Report, pp. 168-171, 21-24 October 2002.

[16] M. Johnson, D. Somasekhar, L. Chiou, and K. Roy, "Leakage Control with Efficient Use of Transistor Stacks in Single Threshold CMOS", iEEE trans. on VLSi Systems, vol. 10, no. I, pp. 1-5, February 2002.

[17] J.c. Park, V. J. Mooney m, and P. Pfeiffenberger, "Sleep Stack Reduction of Leakage Power," Proceeding of the International Workshop on Power and Timing Modeling, Optimization and Simulation, pp. 148-158, September 2004.

[18] Paridhi Athe and S. Dasgupta, "A Comprarative Study of 6T, 8T and 8T Decanano SRAM cell", iEEE Symposium on industrial Electronics and Application (ISIEA 2009), vol.2, pp. 889-894, Kuala Lumpur Malaysia, Oct. 4-6, 2009.

[19] Weijie Cheng, Baolong Zhou, Huarong Zheng, and Yeonbae Chung, "Stack-Transistor Based Differential 8T SRAM Cell for Embedded Memory Application", iEEE International Conference on Electron Devices and Solid State Circuit (EDSSC 2012), pp. 1-2, Bangkok, Dec. 3-5, 2012.

[20] A Feki, B. Allard, d.Turgis, J. Lafont, and L. Ciampolini, "Proposal of a new ultra low leakage lOT sub-threshold SRAM bitcell", international SoC Design Conference (ISOCC 2012), pp. 470-474, , Jeju Island, Nov. 4-7, 2012.