SRAM BASED IN-MEMORY MATRIX VECTOR MULTIPLIER

K.G.Venkata krishna [1], P. Hema naga sai surya kumar 2, S. Meghana 3, A. Reddy prasad reddy [4], G. Muni jayanth [5]

[1] Assistant Professor, Department of Electronics and Communication Engineering, Krishna University College of Engineering and Technology Krishna University, Machilipatnam Andhra Pradesh, India. .
[2] U.G Student, Department of Electronics and Communication Engineering, Krishna University, Machilipatnam, Andhra Pradesh, India.
[3] U.G Student of Department of Electronics and Communication Engineering, Krishna University, Machilipatnam Andhra Pradesh, India.
[4] U.G Student of Department of Electronics and Communication Engineering, Krishna University, Machilipatnam Andhra Pradesh, India.
[5] U.G Student of Department of Electronics and Communication Engineering, Krishna University, Machilipatnam Andhra Pradesh, India.

---***---

Abstract - *The weights stored in the SRAM are turned into proportional voltages using a D/A converter, which is how the SRAM-based matrix-vector multiplier for in-memory computation functions. These voltages are subsequently multiplied by a switched-capacitor stage using an m-bit digital input activation. Finally, charge sharing is used to gather the output voltages associated with the various multiplication outcomes along one column.*

The needed circuit size, calculation time, and power consumption grow linearly with the specified architecture. For the energy usage in switches and capacitors, analytical formulae are provided. Additionally, the effect of manufacturing mismatch on the precision of analogue computing is looked at.

Key Words: **Analog Computation, Hardware Accelerator, In-Memory Computation, SRAM, DRAM**

1.INTRODUCTION

During computations, a lot of data is sent back and forth between the physically distinct memory and processor units of standard Von-Neumann computing systems. It is necessary to reevaluate both the well-established charge-based memory technologies, such as SRAM, DRAM, and Flash, as well as the emerging resistance-based nonvolatile memory technologies in order to get around the limitations of the traditional Von-Neumann-based architectures, which enforce an assertive separation of the processing unit and the memory subsystem.

It is becoming more and more obvious that switching to computing architectures with co-located logic and memory is necessary for application domains like artificial intelligence (AI). IMC, a unique non-Von Neumann computing paradigm, uses the physical characteristics and dynamical state of charging resistance-based memory devices to conduct certain computations directly in the memory. An IMC-based system may be used to accomplish a variety of computing tasks, including logical operations, arithmetic operations, and even certain machine learning activities.

1.1 Motivation

The need for low-power integrated circuits has greatly increased over the past several years as a result of the increasing expansion of battery-operated devices including wireless communication units, portable entertainment devices, and implementable bio-medical chips. SRAM will eventually account for more than 60% of SoCs, predicts the International Technology Roadmap for Semiconductors (ITRS). The problem of consuming power and space is significantly solved when the technology scales by greatly increasing the transistor density in the SRAM units.

1.2 Objective

The in-memory matrix-vector multiplier built on SRAM has as its primary goal a reduction in the amount of time required to complete computations. Performance may be improved and power consumption can be decreased by utilising SRAM technology.

2. LITERATURE SURVEY

2.1 Static Random Access Memory.

In SoCs, embedded SRAMs may take up the bulk of the chip space. Modern scaled-down technologies' increasing process spreads and non-catastrophic defect-related vulnerability to external factors might jeopardise SRAM cells' stability, which is measured by their low Static Noise Margin (SNM). In a cell whose SNM is sufficiently tiny that it might mistakenly flip under the worst operating conditions, a Stability Fault (SF) can occur. The study was done on a thorough SRAM SNM sensitivity analysis and it pinpointed the main causes of poor SNM. A weak Cell Fault Model was presented based on the findings, which may be used in fault simulations to simulate an SRAM cell with a weakened SNM. The SNM of the freshly revised load-less 4T SRAM cell was also given an analytical expression. Several sorts of flaws in the cell's pull-up route may go undetected while reading a 6T SRAM cell with bit lines recharged to VDD. These flaws may result in the SFs. Two fully working SRAM test chips—an asynchronous SRAM (CMOS 0.18 m technology) and a synchronous SRAM (CMOS 0.13 m technology)—are created for the verification of these methods. This approach offers better fault coverage and flexibility than the DRT, shorter test times, and no high-temperature needs. Regular SRAM March Tests have been demonstrated to have a very low detection sensitivity for SRAM cells with possible SFs. The pass/fail threshold's programmability enables tracking of process changes and/or changes to the quality standards without the need for post-silicon design updates.

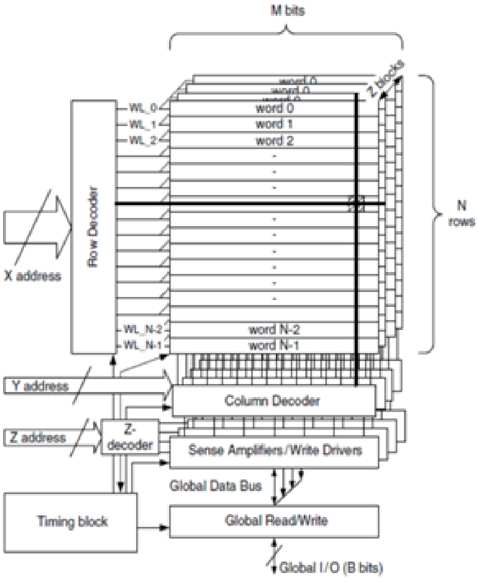


Fig-1: The block diagram for SRAM.

2.2 SRAM Block Structure

An example of the fundamental SRAM block structure is shown in the above graphic. A word line from WL 0-WL N-1 is chosen by a row decoder that is gated by the timing block after decoding the X row address bits. An additional Z-decoder activates the accessed page in the case of an SRAM array with N rows and M bits set up in a page-like fashion.

Word- or bit-oriented memories are also possible. Each address in a bit-oriented memory may access a single bit.

2.3 SRAM Cell

The essential elements of any SRAM used to store binary data are memory cells. Two cross-coupled inverters that create a latch and access transistors make up a standard SRAM cell. Access transistors provide for read-only and write-only cell access as well as cell isolation in the unaccessed state. As long as the cell is powered, an SRAM cell must have non-destructive read access, write capability, and unlimited store (or data retention) duration. Memory cells are organized hierarchically into cores, which may then be further subdivided into blocks and arrays based on the system performance and power needs. A resistive load four-transistor (4T) SRAM cell, a six-transistor (6T) CMOS SRAM cell, and a load-less 4T SRAM cell are three of the most contemporary SRAM cells. A smaller cell increases the amount of bits per unit area and lowers the cost per bit. Because the related capacitances are less with smaller cells, speed and power consumption can be indirectly improved.

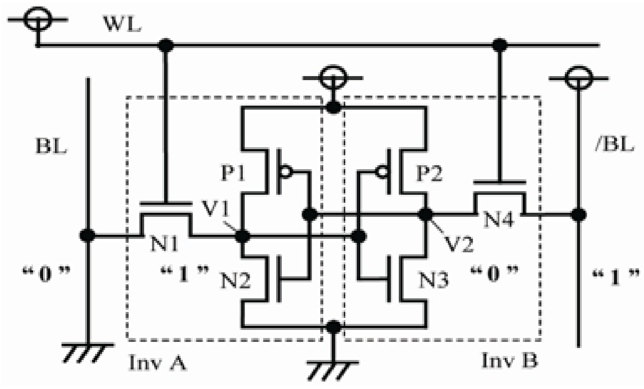


Fig-2: SRAM Cell.

2.4 Designed by 8T-SRAM as ADOT

The schematic for a typical 8T bit-cell may be found here. A decoupled read port is made up of two extra transistors in addition to the well-known 6T-SRAM bit-cell. The write word-line (WWL) must be enabled and the write BLs (WBLs/WBLBs) must be driven to ground or VDD, depending on the bit that has to be saved. The read BL (RBL) must be recharged to VDD and the read WL (RWL) must be activated in order to read a value from the cell. Keep in mind that the source-line (SL) is grounded

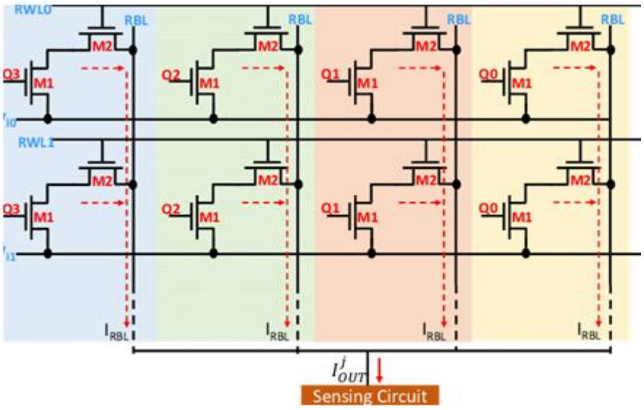


Fig-3: Computing Dot-Product with 4-Bitweight using an 8T-SRAM Memory Array.

3. PROPOSED SYSTEM

Demand for fast speed, low power, and low noise systems is quite strong. Static Random-Access memory (SRAM) can be utilized for several purposes. The dominant matrix-vector operations, according to the idea of in-memory computing for neural network applications, are carried out in the memory itself. The precision of analogue MAC operations is a problem for in-memory computing. By running MAC operations on a regular SRAM, the accuracy barrier is overcome. The initial step in the strategy is to achieve linearly scalable computing accuracy in terms of time, power, and area.

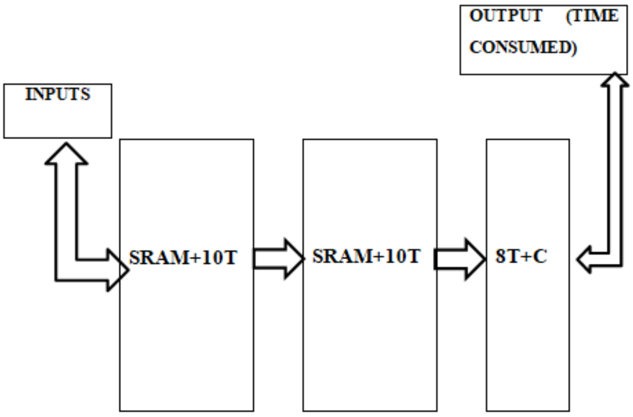


Fig-4: Block diagram of the proposed system.

3.1 Transmission Gate

A transmission gate (TG) is an analogue gate, similar to a relay, that may be controlled by a control signal with nearly any voltage potential to conduct or block current in either direction. It is a CMOS-based switch in which PMOS transmits a strong 1 but a poor 0, and NMOS transmits a powerful 0 but a weak 1. Both NMOS and PMOS function at the same time. Two field-effect transistors (FET) make up a transmission gate, however unlike conventional discrete field-effect transistors, the substrate terminal (bulk) is not internally linked to the source terminal in Fig. 3.2. The drain and source terminals of the two transistors, an n-channel MOSFET and a p-channel MOSFET, are linked together to form a parallel connection. A NOT gate (inverter) links their gate terminals together to produce the control terminal.

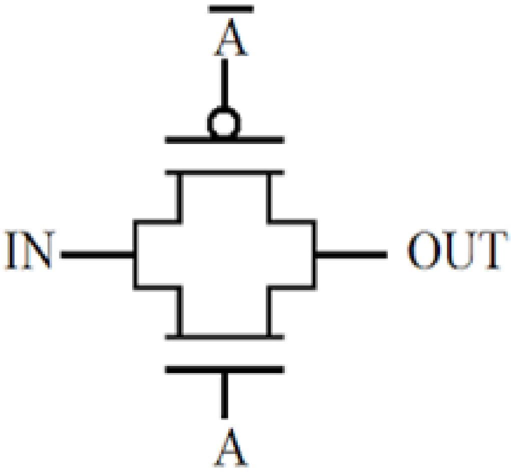


Fig-5: Transmission gate.

3.2 Switched Capacitor

A switched capacitor (SC) is an electrical circuit that carries charges into and out of capacitors in response to the opening and closing of electronic switches. The switches are often controlled by non-overlapping clock signals so that not all switches close at once. Switched-capacitor filters are those that use these components as opposed to exact resistors and rely solely on the ratios between capacitances and the switching frequency. As a result, they are far more appropriate for use in integrated circuits than precisely defined resistors and capacitors, which are more expensive to build.

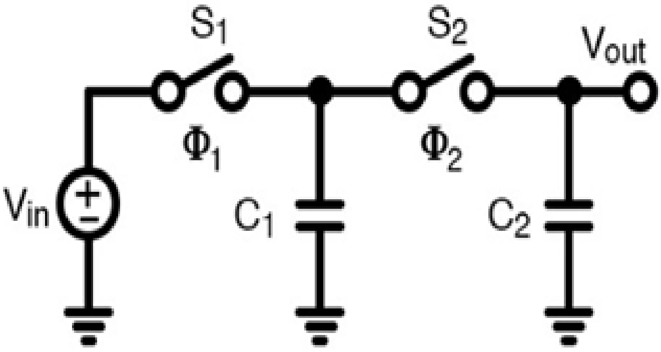


Fig-6: Circuit with switched capacitors.

SC circuits are generally constructed using the complementary CMOS (CMOS) process and implemented utilising metal oxide semiconductor (MOS) technology, including MOS-capacitors and MOS field-effect transistor (MOSFET) switches. Pulse code modulation (PCM) codec-filters, analogue to digital converter (ADC) chips, mixed signal integrated circuits, and PCM digital telephony are a few examples of common uses for MOS SC circuits.

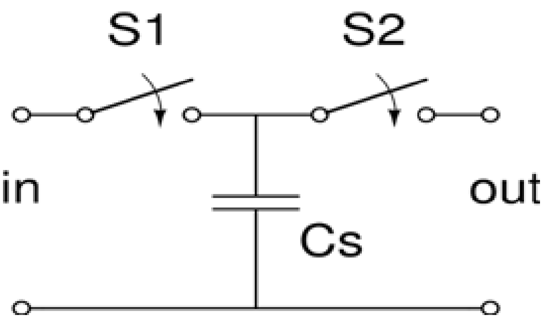


Fig-7: Capacitor resistor with switch.

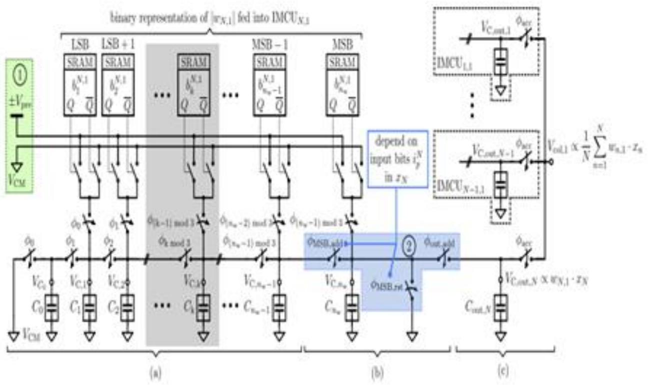


Fig-8: Circuit diagram for a switched leave capacitor.

3.3 Analog Multiplication.

The voltage that is proportionate to the weight must then be analogly multiplied with the input as the following step. Similar to the weight Wn, it is expected that the input Xn is represented in SMR as a nx-bit fixed-point number. So, by performing an XOR operation between the respective signs of the weight (bn sign) and input (i n sign), it is possible to determine the sign of the multiplication result Sn result right away.

$$S^n_{result} = \begin{cases} +1, if\left(b^n_{sign} \pm i^n_{sign} = 0\right) \\ -1, otherwise \end{cases}$$

As a consequence, the Sn result may be used to determine the Vpre recharge voltage. Figure displays an example of a transistor-level implementation of an SRAM-based 3-bit signed IMCU and the related circuit implementation. The recharge voltage selection step and the corresponding circuitry can be skipped in the event of an unsigned multiplication. A multibit fixed-point multiplication of an input Xin by a weight Wn can also be restated as a sum of nx binary products if the distributive law is applied.

$$s^n_{result}.|w_n.x_n| = s^n_{result}.|w_n|.\sum_{p=1}^{n_x}\left(i^n_p.2^{-p}\right)$$

As a result, while processing each bit of the input individually, the multiplication may be done consecutively in nx multiply and add stages. The best way to accomplish this in hardware is to change the control signals on the switches of the MSB capacitor Cnw, which is charged to Vw, n at the ncyc, w-th cycle.

$$\emptyset_{MSB,add} = i^n_p AND\emptyset_{(n_w)} mod3$$

$$\emptyset_{MSB,rst} = \sim i^n_p AND\emptyset_{(n_w)} mod\ 3$$

Despite having many similarities to operational amplifiers, analogue multiplier circuits are far more prone to noise and offset voltage-related issues since these mistakes can compound. Phase-related issues may be very complicated when working with high-frequency transmissions. Wide-range general-purpose analogue multipliers are far more difficult to manufacture than operational amplifiers, and they are frequently made utilizing specialized technologies and laser trimming, just as high-performance amplifiers like instrumentation amplifiers. Since they are quite expensive, they are often only employed in circuits where they are absolutely necessary.

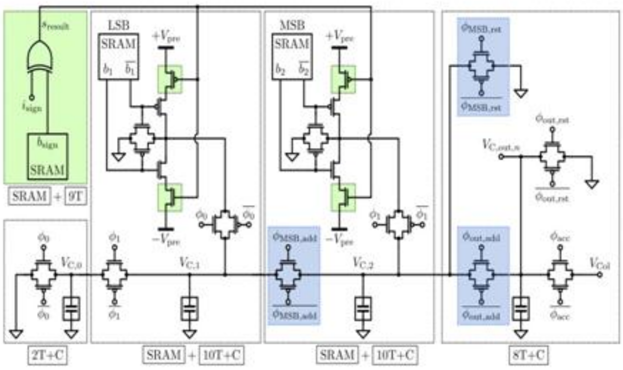


Fig-9: IMCU implementation at the transistor level using SRAM cells.

4. SOFTWARE REQUIREMENTS

A set of tools for designing integrated circuits is called Tanner EDA. With these tools, you may enter schematics, run SPICE simulations, create physical designs (such chip layouts), verify design rules (DRC), and do layout versus schematic (LVS) comparisons.

Fig:-10: The tool's name.

4.1 Design Tools

Three different tools are

- S-edit
- T-SPICE L- edit
- A schematic capture tools
- the SPICE simulation engine integrated with S-edit - the physical design tool

5. RESULTS

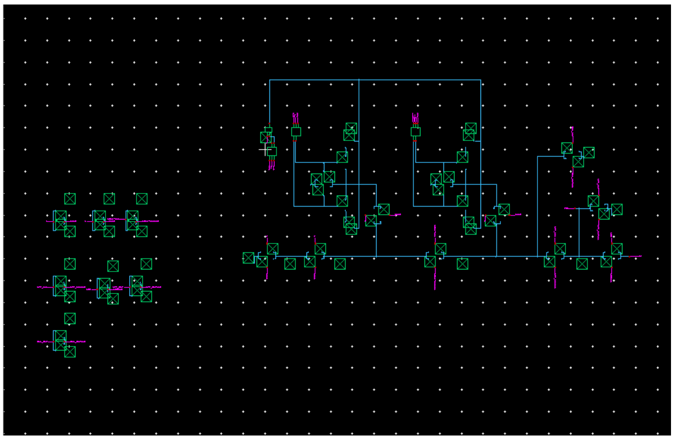


Fig-11: Schematic view

The schematic perspective is depicted in the above diagram. They have revealed An Sram's internal relationships as well as its structural details.

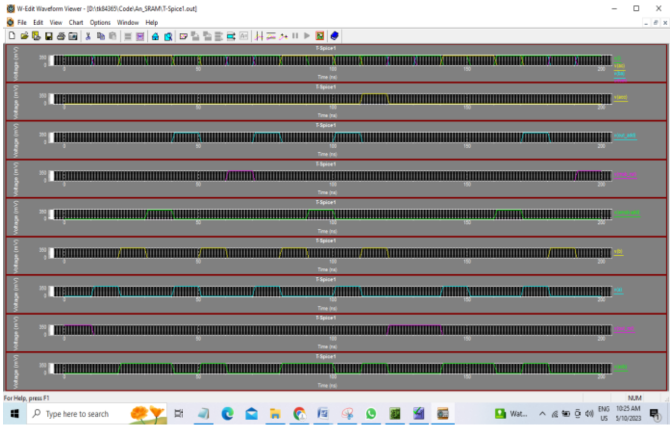
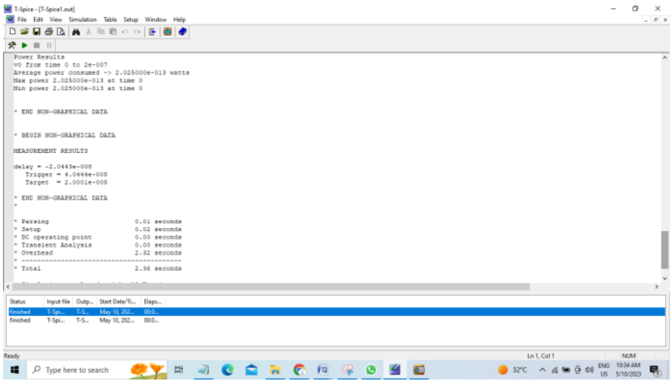


Fig-12: Output waveforms.

The SRAM's waveforms are seen in the above graphic. When we divide a wave into its many parameters.



```
Power Results
v0 from time 0 to 2e-007
Average power consumed -> 2.025000e-013 watts
Max power 2.025000e-013 at time 0
Min power 2.025000e-013 at time 0


* END NON-GRAPHICAL DATA


* BEGIN NON-GRAPHICAL DATA

MEASUREMENT RESULTS

delay = -2.0443e-008
   Trigger = 4.0444e-008
   Target  = 2.0001e-008

* END NON-GRAPHICAL DATA
*

* Parsing                    0.01 seconds
* Setup                      0.02 seconds
* DC operating point         0.00 seconds
* Transient Analysis         0.00 seconds
* Overhead                   2.32 seconds
* -----------------------------------------
* Total                      2.36 seconds
```

Status	Input file	Outp...	Start Date/Ti...	Elaps...
finished	T-Spi...	T-S...	May 10, 202...	00:0...
finished	T-Spi...	T-S...	May 10, 202...	00:0...

Fig-13: Power and Time Results

The effect of weight and input quantization on the maximum power, average power, and multiply operation time. Due to the fact that every extra weight bit requires a new set of capacitors, switches, and clock cycles, delay and power consumption both rise linearly. As a result, their product, which is energy consumption, exhibits a square dependency. Nevertheless, for the input bits nx, the scaling vs energy stays perfectly linear since pipelining simply necessitates three more cycles of operation for the circuit, with no additional hardware needed inside the IMCU.

3. CONCLUSIONS

The idea of in-memory computing for neural network applications has been motivated by the time and energy costs involved with data transportation. This method makes use of specific physical characteristics of memory technologies to conduct the dominant matrix-vector operations in-place, or in the memory itself. Our method is the first to achieve computing accuracy that grows linearly in time, power, and area, despite the fact that there are several SRAM-based matrix-vector multiplication engines in the literature.

The precision of the analogue MAC operations is the fundamental obstacle to in-memory computing. The area and power needed can be decreased by switching to 14 nm technology from 45 nm technology.

The SRAM-based multibit in-memory matrix vector multiplier (IMMVM), which has the potential to dramatically increase the speed and power efficiency of a variety of computing workloads, is a promising technology. Matrix-vector multiplication (MVM) may be carried out by the IMMVM architecture directly in the memory array, obviating the requirement to transfer data back and forth between memory and the processing unit. This can increase overall system effectiveness and lessen the data flow bottleneck.

REFERENCES

[1] 1.B. Keeth and R. J. Baker, DRAM Circuit Design: A Tutorial, 1st ed. Hoboken, 9NJ, USA: Wiley, 2000.

[2] P. F. Ferguson, X. Haurie, and G. C. Temes, “A highly linear low-power 10-bit DAC for GSM,” in Proc. IEEE Custom Integer. Circuits Conf., May 2000, pp. 261–264.

[3] M. Le Gallo et al., “Mixed-precision in-memory computing,” NatureElectron., vol. 1, no. 4, pp. 246–253, Apr. 2018.

[4] W. C. Jeong et al., “True 7 nm platform technology featuring smallest FinFET and smallest SRAM cell by EUV, special constructs and 3rdgeneration single diffusion break,” in Proc. IEEE Symp. VLSI Technol., Jun. 2018.

BIOGRAPHIES

K.G.VENKATA KRISHNA, Assistant Professor, Krishna University College of Engineering and Technology Krishna University, Machilipatnam.

PARASA HEMA NAGA SAI SURYA KUMAR, Student of Department of Electronics and Communication Engineering, Krishna University, Machilipatnam.

SARIKOKKU MEGHANA, Student of Department of Electronics and Communication Engineering, Krishna University, Machilipatnam.

ATURI REDDY PRASAD REDDY, Student of Department of Electronics and Communication Engineering, Krishna University, Machilipatnam.

GODUGUCHINTHA MUNI JAYANTH, Student of Department of Electronics and Communication Engineering, Krishna University, Machilipatnam.