UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Low Matchline Voltage Swing Content-Addressable Memory Cell For Energy-Efficient Search Operations

Mecanismo de Titulación: Tesis en torno a una hipótesis o problema de investigación y su contrastación

Cristhopher Rafael Mosquera Montero

Marco Lanuzza, PhD.
Ramiro Taco, PhD.
Esteban Garzon PhD.
Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título de Magíster en Nanoelectrónica

Quito, febrero de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Low Matchline Voltage Swing Content-Addressable Memory Cell For Energy-Efficient Search Operations

Cristhopher Rafael Mosquera Montero

Nombre del Director del Programa:	Luis Miguel Prócel
Título académico:	Doctor of Philosophy
Director del programa de:	Maestría en Nanoelectrónica

Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor of Philosophy
Decano del Colegio:	Colegio de Ciencias e Ingenierías

Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor of Physics

Quito, febrero de 2025

© DERECHOS DE AUTOR

Nombre del estudiante:	Cristhopher Rafael Mosquera Montero
Código de estudiante:	00331883
C.I.:	0954223046
Lugar y fecha:	Quito, 25 de febrero de 2025

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

AGRADECIMIENTOS

I would like to express my sincere gratitude to my supervisors. I would like to thank Professor Marco Lanuzza for bringing his considerable experience and knowledge to this project. Thanks also to Dr Ramiro Taco who acted as mentor throughout my studies and provide invaluable guidance. I would like to thank Dr Esteban Garzon for sharing his knowledge and expertise on the research. I would also like to extend my gratitude to Dr Benjamin Zambrano who acted as an unofficial mentor both academically and in day-to-day difficulties. I want to thank Director Luis Miguel Procel for giving me the opportunity to achieve my goals and for the trust he has placed in me.

I would also like to give special thanks to my parents for their continuous support and encouragement from the very beginning. Special thanks to my cousin Nataly who was present during all the writing process. Finally, thanks to all the people who were in the process and caused an improvement in my project and in me.

RESUMEN

Las memorias CAM son memorias asociativas que compara la información de entrada con la información almacenada en busca de coincidencias. Las CAM ofrecen la dirección de las palabras donde se almacena el dato buscado o indican si no existe ninguna coincidencia. Estas memorias han ganado gran importancia en campos que requieren operaciones de búsqueda como aprendizaje automático, análisis de datos y secuenciación de ADN. El presente trabajo ofrece una CAM en 65 nm de 256×128 con una matchline de bajo voltaje que reduce un 29% el consumo energético y aumenta un 42% la velocidad de búsqueda. El diseño logra mantener un área similar a las arquitecturas actuales, junto con su bajo consumo es óptimo para aplicaciones que requieren manejar una cantidad masiva de información.

Palabras clave: memorias asociativas, CAM, bajo consumo, búsqueda rápida, low-swing, matchline.

ABSTRACT

CAMs are associative memories that compare the input data with the stored data in search of a match. CAMs return the address of the words where the searched data is stored or indicate if no match exists. These memories have gained great importance in fields that require search operations such as machine learning, data analysis and DNA sequencing. The present work offers a 65 nm 256×128 CAM with a low-voltage matchline that reduces power consumption by 29% and increases search speed by 42%. The design maintains an area similar to current architectures, the area and its low power consumption are optimal for applications that require handling a massive amount of information.

Key words: associative memories, CAM, low power, fast search, low-swing, matchline.

Contents

List of Figures

Chapter 1

Introduction

Rapid advancement in technology demands increasingly specialized and high-performance circuits. These circuits are essential for various tasks, such as addition, multiplication, information storage, and even managing other circuits. As technological demands grow, circuits are being designed to perform specific tasks with greater efficiency. For instance, memory circuits were originally developed solely for data storage. Over time, specialized memory types emerged to meet specific needs. Read-only memory (ROM) was created for situations where data only needed to be read, while static random access memory (SRAM) was designed for applications requiring low access times. On the other hand, dynamic random access memory (DRAM) became popular due to its cost-effectiveness and compact size [3]. Today, search algorithms and increased Internet traffic require circuits that store and compare information.

Content-addressable memories (CAMs) are a type of associative memory that enables fast, massively parallel search operations. This capability has been leveraged in various applications that require intensive search workloads, such as signal processing, approximate search operations, big data analytics, machine learning, and DNA sequencing [4]. A CAM can search and compare data in a single clock cycle. The basic operation of a CAM involves performing a bitwise comparison between its stored data and the input search data. Upon completion of the search, the CAM signals a HIT and provides the address of the matched data [5].

Standard CAM designs are classified into two types: NAND and NOR CAMs. Each type presents trade-offs in terms of performance and energy con-

sumption, both relying on a core CAM cell formed by a 6T-SRAM, but differing in the comparison subcircuit. Although NAND-type CAMs present performance and scalability limitations, they are preferred in power-constrained applications, while NOR-type CAMs are typically used in high-speed applications [6]. The lower performance of NAND-type CAMs has limited their range of applications, whereas the NOR-type has been widely adopted, incorporating several techniques to reduce power consumption.

Various techniques have been applied to NOR-type CAMs to reduce power consumption while maintaining high-speed search operations. At the architectural level, in [7] a hierarchical approach is used to search and compare data through global and local search lines in a segmented array. This technique saves energy by avoiding unnecessary precharge phases, though the need for precise clock signals and flip-flops increases the design's complexity. At the circuit level, in [8] authors have combined the low power consumption of NAND-type CAMs with selectively precharged NOR-type CAM segments. While the segmented wordline reduces capacitance in the *matchline*, and the NAND-type CAM decreases the likelihood of precharging the NOR-type segment, the sequential search operation of the NOR-type CAM significantly diminishes performance. Another effective technique for reducing search delay and power consumption involves a self-referenced sense amplifier connected to the *matchline* [9]. This approach reduces voltage swing and provides high robustness against process variations.

In this work, a low *matchline* voltage swing CAM is proposed to alleviate power consumption and maintain high speed search operation. By a simple modification in the basic CAM cell, voltage swing in the *matchline* is reduced. The technique presents robustness against process, voltage and temperature variations without the power-hungry sense amplifier. Compared to conventional precharge high NOR-type CAM the proposed technique achieves 42% higher speed and 33.8% less energy consumption. Postlayout results demonstrate that the minimum operating voltage is 0.4 V and correct operation is assured for a wide temperature range.

Main objective

- Design and layout of a 256 × 128 low *matchline* voltage swing content addressable memory for energy-efficient search operations

Secondary objectives

- Characterize the core cell for search operations under process, voltage and temperature variations
- Analyze trade offs in terms of search speed, energy consumption and area occupancy
- Compare the technique to a conventional precharge high NOR-type and energy efficient design

The current work is organized as follows: Chapter 2 provides the background, with a description of the conventional CAM and some low-power techniques and methods to increase the performance of the CAM. In Chapter 3, a detailed description of the proposed CAM cell and the architecture of the 128-bit word for a low matchline voltage swing is presented. Chapter 4 shows post-layout comparison results between the conventional NOR type CAM and the proposed technique, and, finally Chapter 5 discusses the results and concludes the work.

Chapter 2

Background

Memory circuits have been a fundamental component of electronics for a long time, serving the essential functions of data storage, writing, and retrieval. Traditional random access memory (RAM) devices store and access data at specific memory locations, known as addresses. During read and write operations, these addresses are assigned, limiting the memory to sequential operations. [10]. However, modern search-based applications demand rapid address comparison and retrieval. To meet this need, Content Addressable Memory (CAM) has emerged as a novel type of memory. Understanding the properties of CAM first requires a brief overview of how RAM functions works.

2.1 CAM vs RAM

Random Access Memories (RAM) are memory arrays designed to store and retrieve data efficiently. These memories support two primary operations: reading and writing, both of which can be performed simultaneously in a single cycle [11]. The inputs to RAM include the data to be stored, control signals that determine whether to read or write, and the specific address of the data. The output is the retrieved data. The speed of data access can vary depending on the memory type. Other constraints of the memory are the area density and energy consumption. As modern applications demand the storage of larger amounts of data with minimal energy usage, RAM continues to perform its tasks effectively. However, emerging applications now require not only rapid access to stored data but also quick retrieval of the associated addresses. To

meet this need, Content Addressable Memories (CAMs) have been developed, which are capable of returning the address of specific data, providing a solution for these advanced requirements.

These memories offer an additional functionality beyond that of standard RAM: they can return the address of specific data. However, this capability requires more complex circuitry and results in higher power consumption compared to RAM. The primary challenge with CAMs is that they need to store the same amount of data as their corresponding RAM but occupy significantly more physical space. Despite this, both types of memory are essential, as they serve different purposes in various applications.

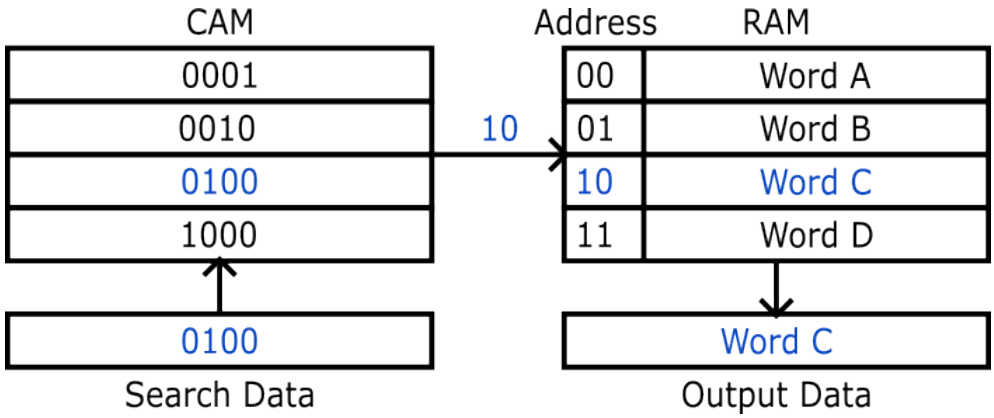


Figure 2.1: CAM and RAM conceptual diagram.

Fig. 2.1 shows how memories work together and the main operation of both, search and read. The diagram can also illustrate some of the main applications. For example, for data encryption, the encryption operation takes the data and with a mathematical function changes the appearance of the text [12]. If the CAM data is taken as the input text, and the RAM stored data and the address is taken as the encryption function, the search data will be returned as encrypted data. The IP addressing operation is similar, and the same applies for the look-up-tables and search based algorithms. Once the importance of the CAMs and their applications is justified, it is important to know how they really work from the array to the individual core cell.

2.2 CAM architecture

A content-addressable memory is organized as an array of $m \times n$-bit words. Fig. 2.2 illustrates a CAM array of m words, with each output (ML_n) connected to the encoder. The CAM receives the Search word as input and performs a bit-wise parallel search across the entire array. After the search operation, each ML_n indicates a HIT or a $MISS$ to the encoder which then outputs the address of the matched word [13].

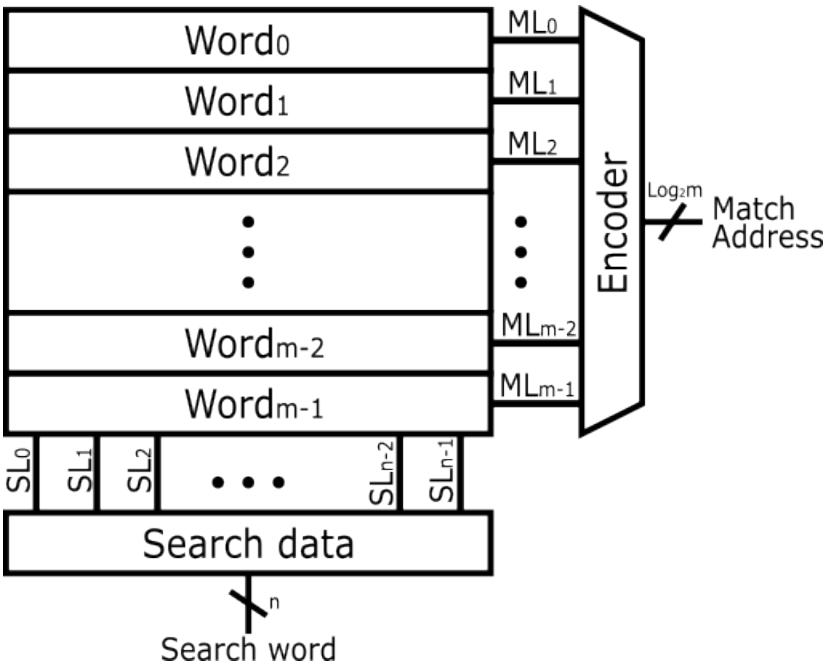


Figure 2.2: Block diagram of content addressable memory of m words of n bits.

As an input the n-bit search data is driven to each word in the array through the Search Lines (SL_n and $\overline{SL_n}$). In the CAM word the corresponding $n^{th} - bit$ of the SL is compared to a previously stored $n^{th} - bit$ in the word. A single word in the array is formed by $n - bit$ CAM cells able to store a bit in a memory element and compare its content (i.e XOR) to the SLs [14].

According to [1] the number of bits in a word depends on the application (i.e from 8 bits for machine learning up to 144 bits for IP address networking) and the number of words range from 128 up to $32K$ divided in banks with dedicated access circuitry.

Fig. 2.3 depicts a conventional segmented NOR-type word structure. The word is composed of n CAM cells grouped in s segments. Each segment

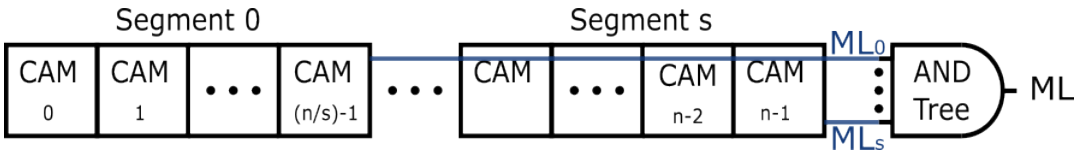


Figure 2.3: Structure of n bits word divided in s segments.

output is connected to a logic circuit, such as the AND tree, to obtain the final result of the comparison of the whole word.

2.3 NOR-type CAM cell

A conventional NOR-type CAM cell is formed by a memory element and a comparison circuit [15]. The memory element requires high speed access and robust bit storage. Fig. 2.4 depicts the schematic of a conventional 10-T NOR-type CAM. The memory element is a 6-T Static Random Access Memory and 4-T comparison circuit. The figure omits the two access transistors which are part of the typical SRAM. A conventional CAM is not able to *write*, *read* and *search* in parallel thus bit lines and search lines are shared in the cell.

Transistors MI1-MI4 compose the inverters that store the data. D and $\overline{D}$ are the stored bit and its complementary. These inverters are connected by the output to the input of the other. It is a loop structure that stores a bit and its complementary value with only 4 transistors. The inverters connected in loop offers high stability in the moment of reading and comparing the data, ensuring that the D value is not inverted or lost.

M1-M4 are transistors that compare the search and the stored data. As can be seen the transistors are nmos and they make the pull-down of a XNOR gate. The inputs are the stored bit, the search bit and their complements. The pull down XNOR structure is connected to the ground as in a normal gate and where the output should be, the ML is connected. It is appreciable that the order of the inputs in the path do not change the logic, the SL and $\overline{SL}$ can be connected to M1 and M2, and the data to M3 and M4. But this configuration where transistors with stored data are closer to the ML was selected to avoid charge sharing issues. Since stored data is stable during search operations the transistors M1 and M2 will be on the same state along all the process.

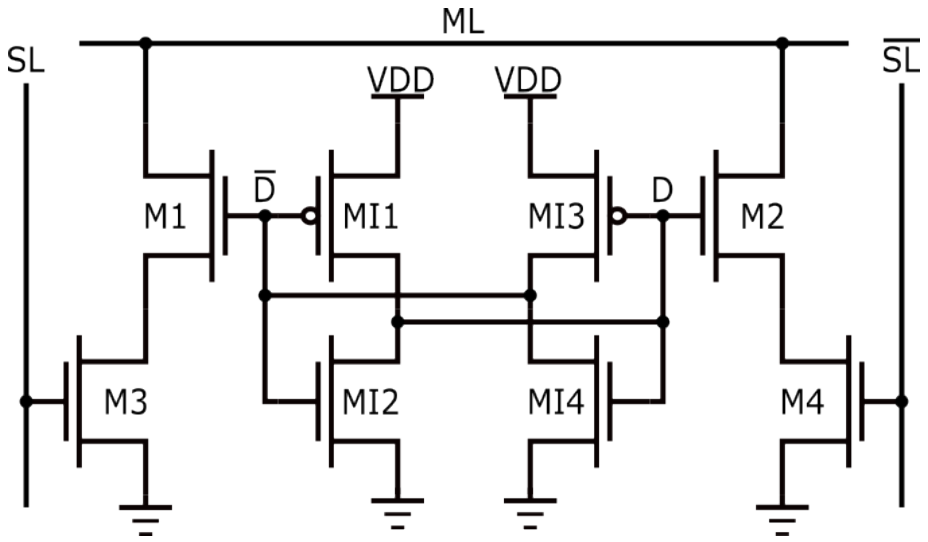


Figure 2.4: NOR cell schematic.

Consequently one of the transistors will be on and the other off connecting the ML to the internal node and charging also this node.

The charge and discharge of the ML is the indicator of a match or mismatch. If there is a match on the cell both paths (M1-M3 and M3-M4) are on open circuit. If there is a miss, some of the path makes a short circuit between ML and ground. Therefore, to get a match in the word, all the bit cells of the word must compute a match and leave the ML as a floating node. Following this logic, a properly connection of the ML can determine if there is a match or not.

2.4 Matchline connection and structure

To assemble a full word circuit with the NOR-type CAMs all the cells are connected in parallel, Fig. 2.5. All cells that constitute a word share the same ML same as the WL, and, each cell has their own SL and $\overline{SL}$. The search lines are shared with the cells of the other words. It is expected that in the array one word is under the previous one and the SLs and $\overline{SL}$s run vertically through the entire memory. In the case of the ML, that goes in a horizontal position, it is also connected to a pmos transistor M_{EN} controlled by an enable signal that also controls the search steps.

The search operation can be divided into three stages. First the pre-discharge of the search lines. The objective of this step is to leave the ML

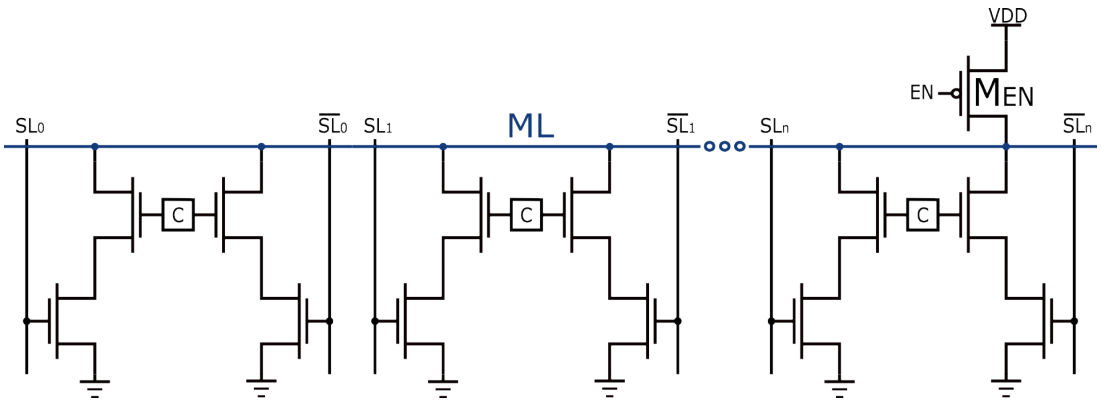
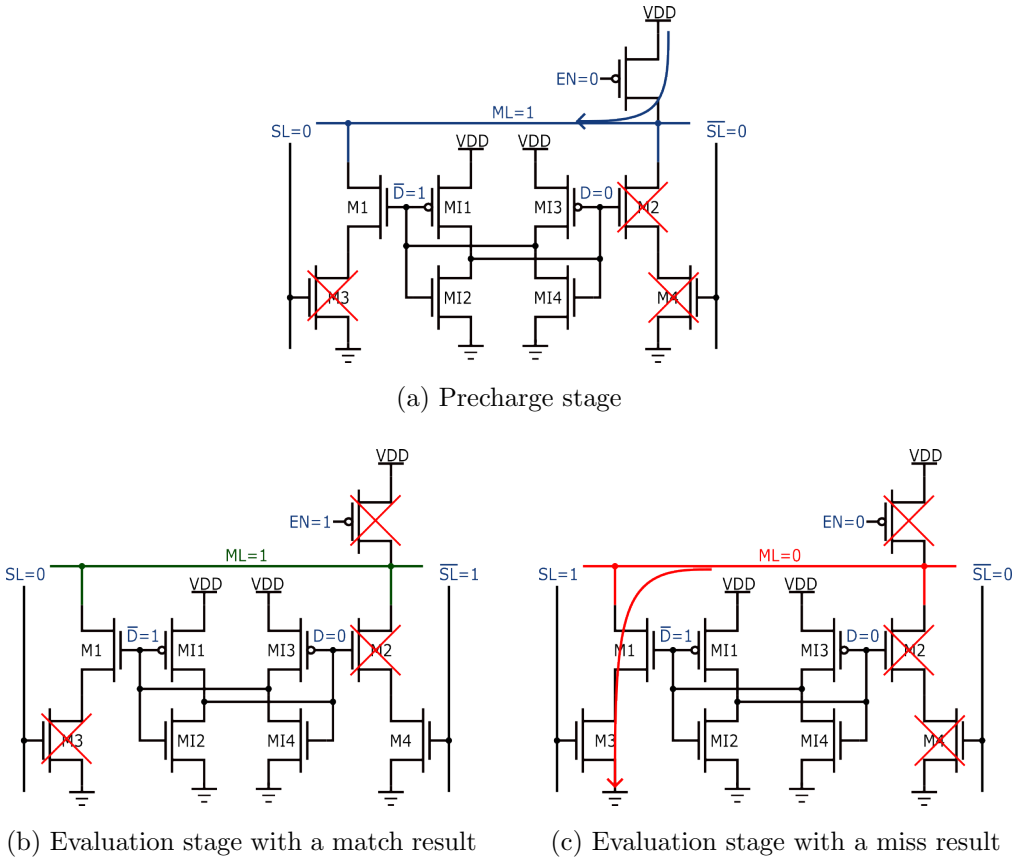


Figure 2.5: NOR-type match line scheme.

floating. Since the NOR-type CAM uses nmos to connect the ML to the ground, the SLs and $\overline{SL}$s should be discharged to turn off the transistors and have an open circuit. The second step is to pre-charge the ML, giving an initial value to allow to determine if there is a match or not. The M5 transistor is turned on with the enable signal. This change of the transistor state connects the ML with the supply voltage and charges the line to V_{DD}. Finally, the M5 transistor is turned off and the search data is broadcast into the SLs and $\overline{SL}$s. If the word matches with the input data all the paths to the ground will remain in an open circuit and the ML value stays on V_{DD}. In the other case, at least one bit mismatch will connect ML to the ground and discharge it, using the 0 value to indicate that the word is different.

These three steps can be reduced to two. The objective of the first two steps is to give the ML an initial value to keep if there is a match. Both steps can be done at the same time. If the enable signal controls the broadcast of the values into the SLs, $\overline{SL}$s and the precharge transistor, the search lines can be discharge, leaving the ML floating, while the M_{EN} charges the ML to high. Now the enable signal can be used like a clock signal and it is said that the search operation is performed in one clock cycle. The first half corresponds to the first two steps and it is called the precharge stage [16]. The second half is the third step, that is the evaluation of the data. Fig. 2.6 offers a graphically explanation of what happened in the bit cell and in the matchline in both steps. In the precharge stage, the sharing issue problem can also be observed. M1 is on and connecting the ML with the internal node between M1 and M3, during the precharge operation this node will also be charged.



(a) Precharge stage

(b) Evaluation stage with a match result

(c) Evaluation stage with a miss result

Figure 2.6: Search steps.

As was explained the word can be composed by one ML of n bits. However having just one big ML has a major drawback. The issue of having one ML for each word is the large capacitance and higher delay [17]. To decrease the search time the word can be segmented. Dividing the word in 2, generates two ML with $\frac{n}{2}$ bits loading it. The capacitance theoretically will be divided by half reducing the time to charge and discharge the line. Nevertheless, reducing the number of bits per ML has a cost. The main trade-off to divide the word into segments is the area. Having many MLs requires an extra logic circuitry to determine if the word is matching or not. Each segment operates individually in the same way as the process previously described.

AND tree is an example of a logic circuit that can determine the match or mismatch. If all the segments return a match, the word is equal to the search one. In the other case at least one segment returns a mismatch, so the word is different. Fig. 2.7 shows the AND tree schematic for a word segmented in 2, 4 and 8 parts. Is is obvious that more segments mean more circuitry and

more area. Also the complexity of the CAM increases since this logic should be in every word and the gates must be fitted in the array, whilst the designer have to consider the space to locate matchlines that must pass through other segments until reach their respective gate. Since this is a digital circuit the energy and delay added of these gates to the CAM is minimal compared with the saved by the reduction of the ML capacitance.

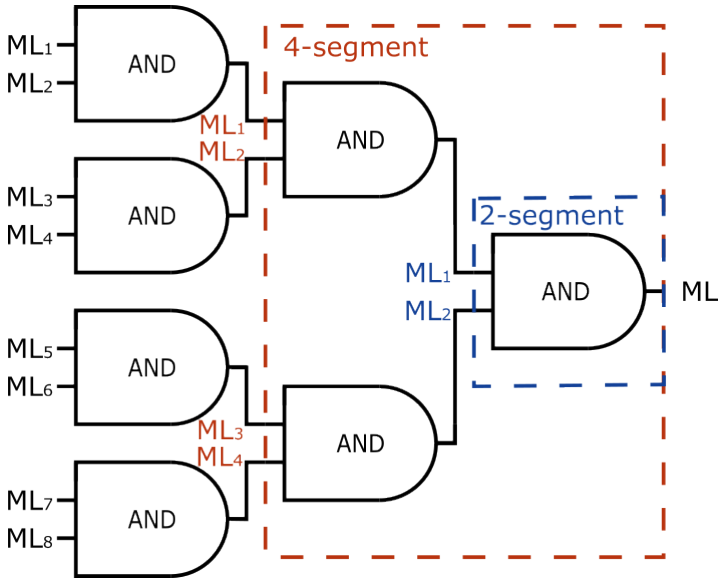


Figure 2.7: AND tree schematic.

This modification of the structure is just one of many that can be found. In order to improve the performance and decrease the energy consumption the structure of the cell and architecture can be modified. Some alternatives to solve it is to use a low-swing schemes, pipelining the search path and employing hierarchical lines.These approaches can reduce the power consumption and the search time.

2.5 Low-Swing Scheme

Low-swing technique consists of reducing voltage range employed by a circuit or some part of it in order to modify the performance of it. Usually the voltage level of a heavily load line is reduced to decrease the power. The reduction is not made randomly, it have to be into the circuit operational range and should not compromise the speed. Low-swing can be applied into the ML and SL which are lines with a bigger capacitance load [18]. Considering the low swing

approach into the ML it can be characterized by the following power equation

$$P_{ML} = C_{ML} V_{DD} V_{MLswing} f \tag{2.1}$$

which indicates the power employed in charging the ML. C_{ML} is the capacitance of the ML, V_{DD} is the nominal voltage supply, $V_{MLswing}$ is the voltage swing of the matchline and f is the operation frequency. Power and voltage swing are directly proportional, therefore a lower voltage swing reduces the consumption. This solution saves a lot of energy in exchange for area and complexity. Most of the currently works elevate the minimum value of the ML to reduce the voltage swing and generates more circuitry to sense it. This sense circuit of the ML like in [19] and [20] also requires a circuit to generate a voltage reference . Those circuits increase the area occupancy and this problem is more noticeable when many ML are used for each word.

2.6 Pipelining and hierarchy

The principles of the mentioned schemes and configurations are applied in the conventional cell design and in the proposed circuit. However there are other techniques that can save energy or improve the performance and are suitable with the conventional CAM and the proposed. On the final part of the work it is discussed that these techniques can work better or worse for the circuits, therefore it is convenient to explain the principles of these techniques.

2.6.1 Pipelined scheme

Pipeline is a technique that aims to increase the frequency of operation in exchange for area. It consists of adding a register into the critical path. More register means less logic and lower delay. To apply this technique to the memory, instead of reducing the logic between the registers, the words are segmented [7]. It requires two or more MLs for each word. Fig. 2.8 shows a 5 segment word. The operation is simple using as an example the graphic where there are 5 register, so making one comparison will take 5 cycles. Although it seems like the search time is longer, while the second cycle compares the second segment, the first segment is making the comparison with the next search data.

When the last comparison is making the other 4 segments are making another 4 comparisons. After each cycle, the CAM returns a comparison, therefore the search operation is effectively using just one cycle. Also, due to the reduction of the word into segments, the frequency of the clock can be bigger. The comparison of one segment after another also offers energy savings.

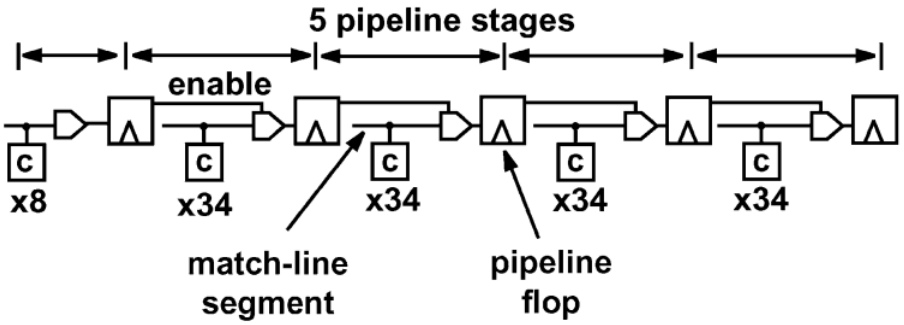


Figure 2.8: Pipelined match-line, extracted from [1].

The word has a match if all of the bits are the same as the search word, thus, all bits are compared. But if some segment returns a mismatch, comparing the rest of the segments is not needed. Since the comparison of segments is made one after another, if the previous segment returns a miss, a flag can shut down the next comparison and notify the next stage that the word is different. E.g. if in the Fig. 2.8 the first segment matches, the second segment performs the comparison of this segment in the next cycle. If it returns a miss, a flag will turn off the comparison of the third segment for the next cycle, and the same will happen for the fourth and fifth segments, saving the energy of these search operations [7].

2.6.2 Hierarchical scheme

As it was said ML is not the unique line with a heavily load in the CAM's memories. SL and SLs also produces a highly captive energy consumption but those lines instead of depending on word size depend on the number of words. One solution is also to apply a low-swing approach, but it strongly depends on the core cell. A more suitable solution to improve the performance is using a hierarchical scheme.

These schemes consist of dividing the lines into global and local lines.

Local lines are directly connected to the bit cells and have a reduced number of bit cells compared to one SL scheme. These local lines are charged or discharged by a gate that is controlled by the global line and an optional enable signal. The global line is connected to all the gates of the local lines of its respective bit. The global line drives the search value into the local lines who drives it into the bit-cells. The advantage of this method is the reduction of the delay. The diagram can be seen on Fig. 2.9. This scheme shows the capacitive load of each SL is reduced and therefore the data is quickly distributed to all the cells. But this technique can be combined to also reduce the energy consumption.

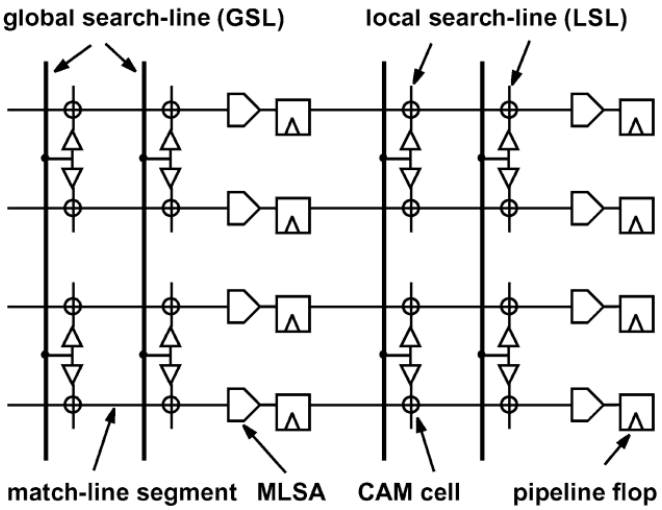


Figure 2.9: Hierarchical technique scheme for CAM, image extracted from [2].

One example is using a low-swing approach in the global lines. The reduced load of the global lines allows the use of a lower voltage level that can be sensed by the local lines gate. This reduce the energy needed and keeps the performance using a full swing in local lines like in [7]. This can also be improved by an enable signal that controls if the local line drives the search value or not. A precomputing of a mismatch can advice that some segments do not require making a comparison, therefore the local lines do not need to broadcast the search data. The pipelined scheme advices if there is a match or not, and shuts down the next segment, smd and the same signal can shut down the local lines that do not require a comparison. Definitively, hierarchical schemes can be combined with the other mentioned techniques to improve the performance of the CAM.

Chapter 3

Proposed bit-cell

As mentioned above the aim of the project is to design a circuit level solution to improve CAM energy usage without sacrificing much area and search time. To achieve the proposed, bit-cell aims to reduce the ML capacitance modifying the conventional structure.

3.1 CAM

The proposed CAM is also made from a SRAM, but adding a XOR transmission gate. Fig. 3.1 shows the schematic of the CAM, where M1-M4 performs the XOR operation, the access transistors are ommited. To improve the search time, nmos and pmos transistors are used on the XOR logic. The gate of those transistors is connected to the stored data instead of the SL and $\overline{SL}$. This avoids connecting the D and $\overline{D}$ to the drain or source of the transistors which can affect the stability of the cell. It also maintains the M1-M4 transistors stable during the search operation. Finally, the XOR output control is a nmos (M5) transistor which is connected to the power supply and the match line.

The proposed CAM cell connects only one transistor to ML for each bit. This structure generates a ML with less capacitance than conventional CAM and decreases the energy consumption since the power is directly proportional to the capacitance. Additionally, the transistor is an nmos that connects the ML to V_{DD}. This transistor allows an low-swing approach for the ML. The cell is an Low-swing CAM (LS-CAM).

As explained earlier, the logic to compare the data is an XOR logic

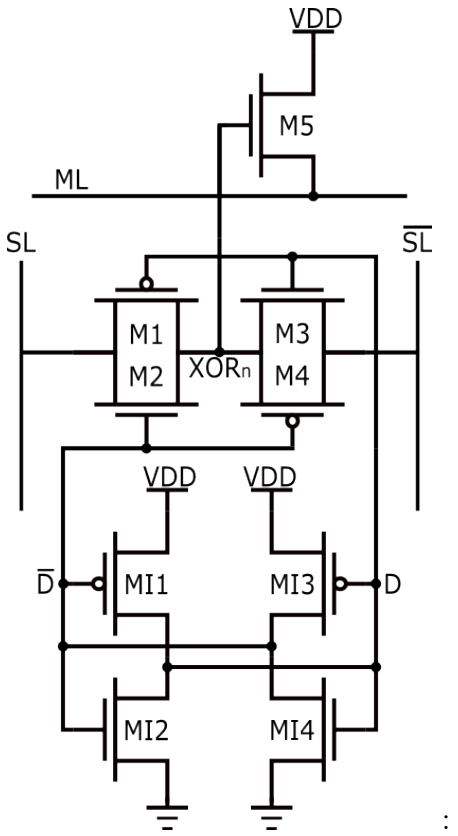


Figure 3.1: Proposed core CAM schematic.

controlling an nmos transistor. The operation is given in the table 3.1. The first case represents the pre-discharge of the SLs and $\overline{SL}$s to disconnect the ML from the V_{DD}. The other four constitute the normal true table of the XOR. The nmos is on off when the comparison returns a match and connects the ML to v_{DD} when the bit is different.

D	$\overline{D}$	SL	$\overline{SL}$	XOR	$M5$
0	1	0	0	0	Off
0	1	0	1	0	Off
0	1	1	0	1	On
1	0	0	1	1	On
1	0	1	0	0	Off

Table 3.1: True table of XOR and M5 behavior.

3.2 Matchline scheme

The LS-CAM is connected in a similar way to the NOR-type CAM. All bit cells of the same word share the same ML node. So all of them are connected in parallel. The main difference is the precharge, for the LS-CAM an nmos transistor controlled by an enable or clock signal is connected between the ML and ground nodes. It also operates with a precharge and an evaluation stage.

The proposed circuit pre-discharges the ML to zero while the SLs and $\overline{SL}$s keeps all the upper nmos transistor off, as shown in the first case of the table 3.1. After the ML reaches the lowest voltage level, the search data is driven into the SLs and $\overline{SL}$s, and the comparison is performed. If there is a match the ML keeps on low, and in the other case, at least one of the nmos transistors of the bit-cells is turning on connecting the ML to VDD. Since nmos is passing a high value it will be a weak high value, $ML = V_{DD} - V_T$. With this structure, a low-swing ML is obtained. Now, a circuit is needed to sense the correct value of the ML.

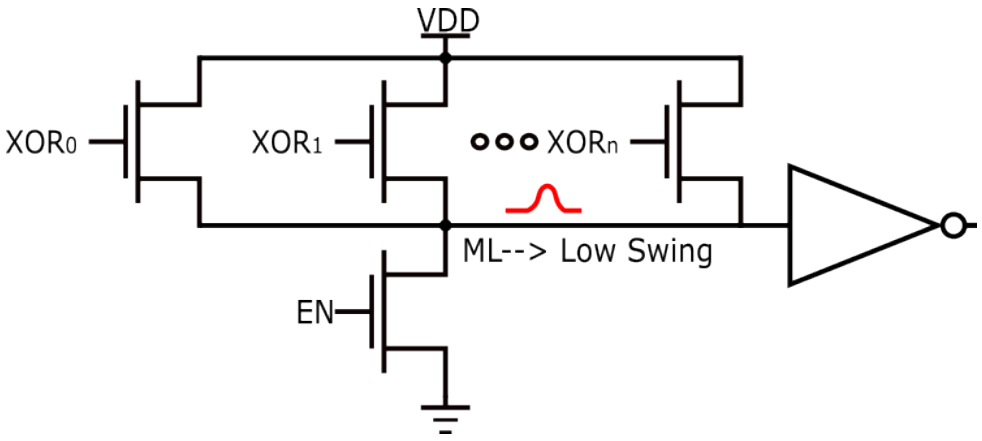


Figure 3.2: LS-CAM matchline scheme.

To sense the low-swing ML, a gate that inverts the value to have the same logic as the NOR-type CAM, 1 for a match result and 0 for mismatch, is needed. Also, it requires the weak high value to be read as a normal high value and take priority over the falling output transition to have a faster evaluation. A LO-skew inverter can execute those tasks perfectly. This inverter is an inverter with a voltage transfer curve shifted to the right, due to the larger size of the nmos compared to the pmos. This circuit has a lower area occupancy, low energy consumption and satisfactory speed.

The proposed ML scheme reduces the capacitance and the voltage swing of the ML, which means a reduction on energy consumption and search time. It does not require a voltage reference or a complex sensing circuit and it can be divided into segments without a large penalty area caused for the sensing circuit. With all these considerations, both circuits were designed to characterize and compare the results.

3.3 Design parameters

The design of both circuits were made with minimum sizes for the comparison transistors (M1-M4) and the 6-T SRAM were designed with the same sizes and transistors for a fair comparison. The dimensions of the transistors for NOR-type CAM are on table 3.2 and for LS-CAM on table 3.3.

Transistor	$w[nm]/l[nm]$	**Type**
M1	200/60	lvt
M2	200/60	lvt
M3	200/60	lvt
M4	200/60	lvt
MI1	120/60	hvt
MI2	120/60	hvt
MI3	120/60	hvt
MI4	120/60	hvt
Access	120/60	hvt

Table 3.2: NOR-type CAM transistors dimensions.

The M1-M4 transistors change their state during the search operation, and to achieve this, the faster operation high voltage threshold (hvt) transistor of $w = 200nm$ were used. For the rest of the transistors, low voltage threshold (lvt) transistors were used to minimize the impact of their consumption in the results.

The comparison transistors, M1-M4, for the LS-CAM are hvt. Since these transistors are on the same state during all the search operation they do not require a low threshold to switch. Additionally, $w = 120nm$ was selected

Transistor	$w[nm]/l[nm]$	**Type**
M1	120/60	hvt
M2	120/60	hvt
M3	120/60	hvt
M4	120/60	hvt
M5	120/60	lvt
MI1	120/60	hvt
MI2	120/60	hvt
MI3	120/60	hvt
MI4	120/60	hvt
Access	120/60	hvt

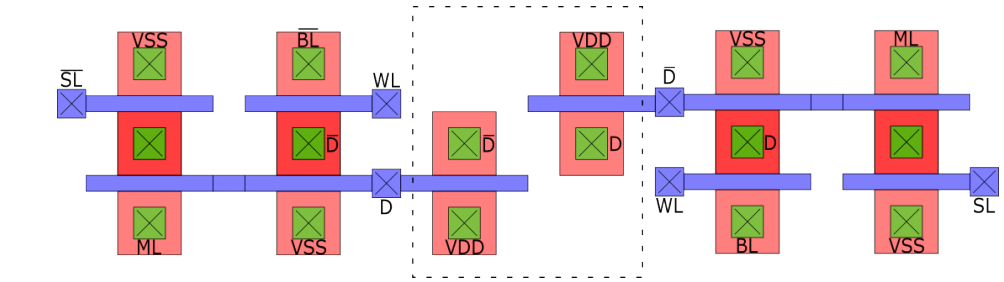
Table 3.3: LS-CAM transistors dimensions.

since two transistors pass the value. Finally the M5 transistor switches on in case of mismatch so an lvt transistor was selected to achieve high search frequency. The rest of the transistors MI1-MI4 and access transistors are designed equal to the NOR-type.

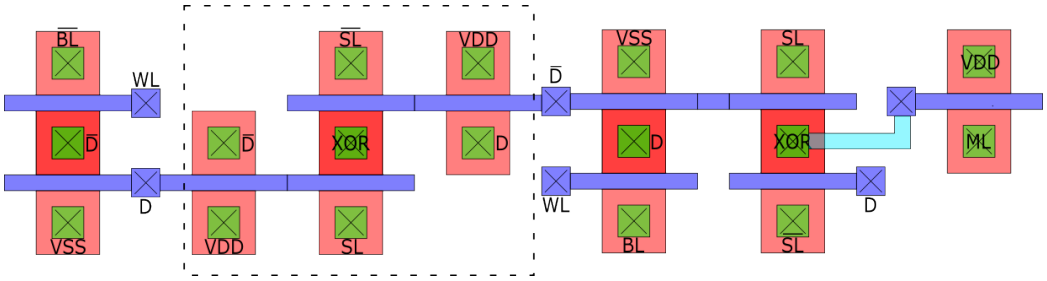
Using the schematic and the transistors dimensions, a diagram stick was performed to reduce the area occupancy and facilitate the connections between cells. Fig. 3.3 shows the stick diagram for both cells. It is expected that both cells have the same height. The width is larger for the LS-CAM, however to match the extra transistor, the layout will be inverted and matched, reducing the extra area. With this diagram, it is expected to use metal 1 for all the possible connections, metal 2 for vertical connections and metal 3 for the horizontal ones.

3.3.1 LO-skew inverter and precharge transistor

The inverter has to read the weak high value of the low-swing ML. To achieve a properly sensing of the value, the maximum ML was measured, and it reached a value of around $650mV$. To have a correct sensing of the value, the vtc was shifted to $300mV$. Finally, for the pmos, a hvt transistor with $w = 120nm$ and $l = 60nm$ was used, and for the nmos, an lvt transistor with $w = 4 \times 240nm$



(a) NOR-type CAM



(b) LS-CAM

Figure 3.3: Stick diagrams. Discontinuous line enclose pmos transistors, which require an NWELL layer.

and $l = 60nm$ was used. This generates a stronger pull-down and offers an LO-skew inverter. Fig. 3.4 shows the VTC of the inverter, that is shifted to the left side, near to the $303mV$ when the output is $500mV$.

The VTC indicates that the inverter reads a high value from 0.4 V with a $V_{DD} = 1$ perfectly. It rapidly returns a zero in case of mismatch and reads the maximum value of the ML, $V_{DD} - V_T$. This inverter is for the LS-CAM and it is necessary in every segment. The NOR-type CAM returns a strong high and low value, so it can be easily read and does not require a low-swing sensing circuit.

Finally, in both circuits, the ML requires a transistor to charge or discharge it. NOR-type CAM employs a pmos transistor for a 64-bit ML, where $w = 15 \times 200nm$ gets the best time results. LS-CAM uses a nmos which is stronger than pmos so the transistor has a $w = 5 \times 200nm$. This larger width compensates the inverter area of each ML. Additionally, since LS-CAM uses an nmos transistor it does not need an NWELL layer, reducing the space of each segment.

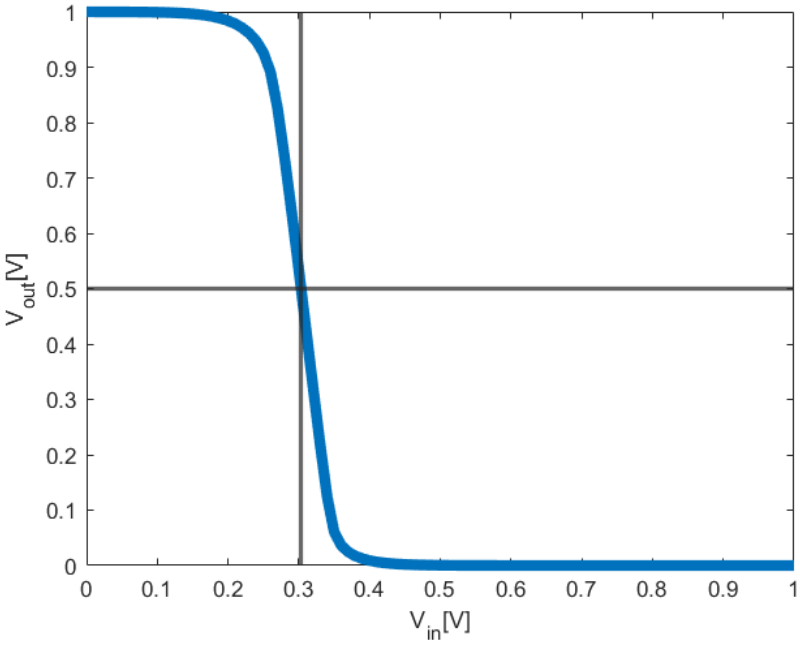


Figure 3.4: Voltage transfer curve of the LO-skew inverter

Chapter 4

Results

The simulations to characterize the cell were performed by the Virtuoso software with a commercial node technology, 65nm of TSMC. The objective parameters are delay and energy, and for voltage, temperature and corner variations. Also, the area is taken into account in the designed layouts. To have a reference values, the NOR-type CAM and the LS-CAM were simulated under the same conditions and with the same peripheral circuitry.

A 128-bit word was assembled with two segments of 64 bits. The SLs and $\overline{SL}$s nodes are loaded with the extracted capacitance of the layout times 255, emulating the load of a 256-word CAM. AND gates pre-discharge the SLs and $\overline{SL}$s, and load the search data. These gates are controlled by the enable or clock signal and have the same dimensions and number of gates for both circuits.

The enable or clock signal also controls the pre-charge transistor of the NOR CAM and the pre-discharge transistor of the proposed CAM. Therefore, for the proposed CAM the enable signal requires an inverter. Finally an extra logic circuitry is necessary to compute the match of the word by the output of the segments. To calculate if the word is equal to the search data an AND logic is performed between the ML of the two segments. If both segments have the ML on a high value (match) the word is the same, and if one or both ML are low (mismatch) the word is different. The inputs of these AND gates are the ML for the NOR-type CAM and the output of the LO-skew inverters of each ML for the LS-CAM.

To get the search time both memories perform a mismatch in one

segment and a match in the other. To produce a mismatch, only one bit is different, it is the worst case for the delay. Two times are measured, the precharge and the evaluation time. The precharge time is measured as the time it takes for the ML to precharge from the time the enable signal initiates the precharge. The evaluation time is measured as the time that the ML takes to indicate the missmatch from the time the enable signal initiates the evaluation. The sum of both times is the search time. To get energy, one segment performs a match and the other a mismatch, but in this case, the miss is generated by 50% of the bits of the segment. The data on the word was loaded with an initial condition file and the inputs were made with a vector file. The power of all the circuitry was integrating to get the energy and divided by the number of bits to present the energy per operation of each bit-cell.

4.1 Layout and Area

The final layout for the NOR-type CAM and the LS-CAM are in Fig. 4.1 and 4.2 respectively. The layouts were designed following the stick diagrams and employing the minimum distance rules to minimize the area occupancy. The SL, $\overline{SL}$, VDD and VSS connections are in vertical with a metal 2, the yellow one. The write line and the matchline are in horizontal with a metal 3, on green. This means that both cells just employ from metal 1 to metal 3,so the rest of metal layers are available to make power connections. Moreover, both cells allow us to use two additional metal 3 rails without generating short circuits and complying with the distance rules. The extra metal layers do not reduce the effective area of the cells and are optimal to employ more segments and to drive more match lines.

To get the effective area, the cells should be connected in the array and measure from the beginning of one cell to the beginning of the next. The array connections are necessary to know the minimum distance that can be between bit cells. For the NOR-type, one cell is located next to the other and the minimum distance is given by the POLY and DIFFUSION layers. But in the LS-CAM, due to the extra transistor, the cell should be inverted and connected while matching the transistor of one cell to the next. A point

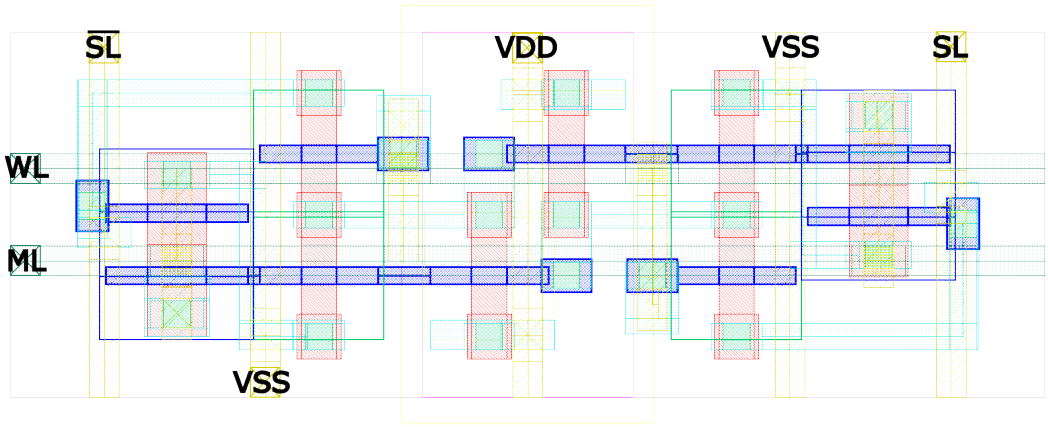


Figure 4.1: Layout NOR-type CAM.

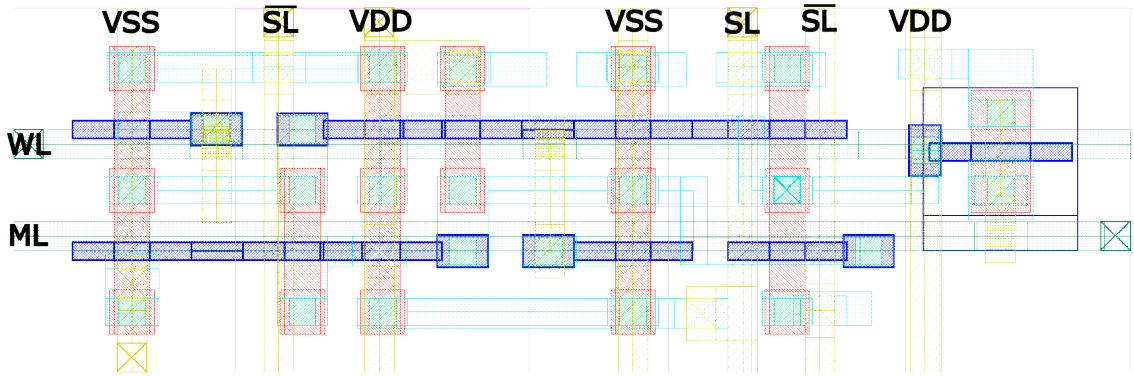


Figure 4.2: Layout LS-CAM.

symmetric layout with the center on the ML transistor's connection is formed by two bit cells. In this case, the measurement is made by each two cells and divided by two to get the real efficient area of the cell. Fig. 4.3 and 4.4 show the memory array made to measure the area of both circuits. The discontinue line indicates the two cells that are matched for the LS-CAM array connection.

The area of the NOR-type cell is $1.08\mu m \times 3.17\mu m$, while the area of two proposed cells is $1.08\mu m \times 6.38\mu m$. This means that the efficient area of one cell is $1.08\mu m \times 3.19\mu m$. The LS-CAM is just 0.63% bigger than the NOR-type CAM.

Finally the parasitic extraction tool offers an approximate of the capacitance load present on the lines. This is just an approximation of the parasitic

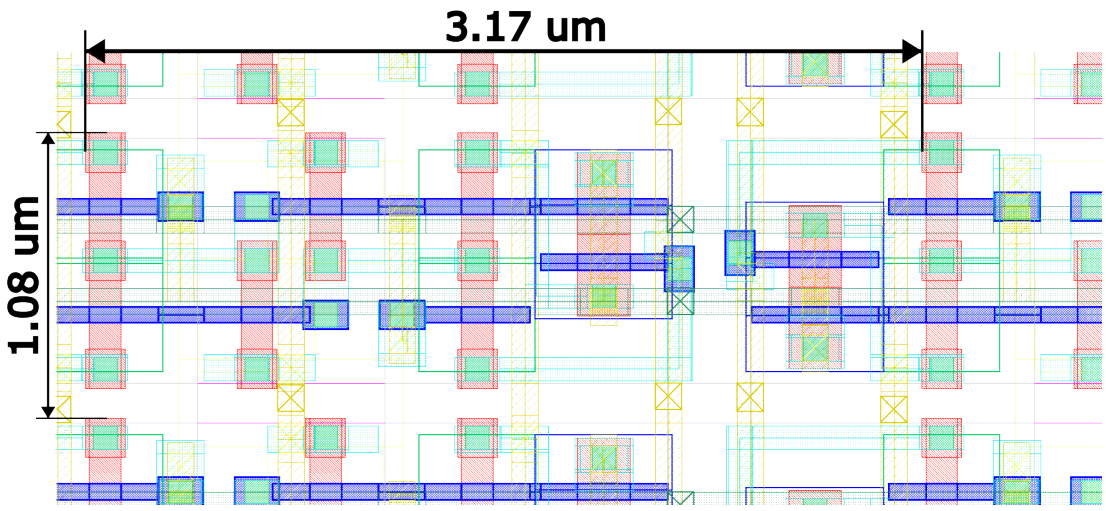


Figure 4.3: NOR-type CAM array.

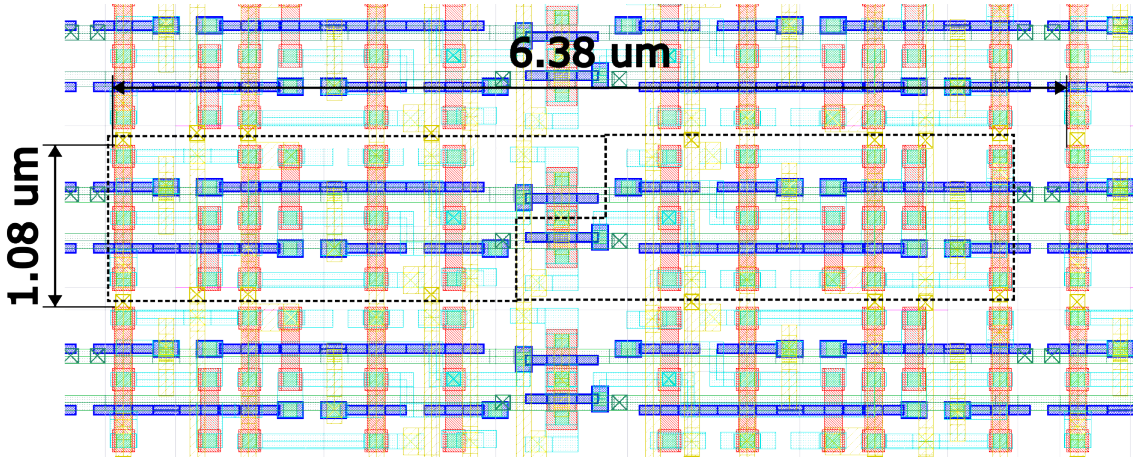


Figure 4.4: LS-CAM array.

capacitance between the critical lines and the ground node. For the NOR-type CAM, each cell adds $270aF$ to the SLs and $344aF$ to the ML, while the LS-CAM adds $177aF$ to the SL, $218aF$ to the $\overline{SL}$ and 161 to the ML. The LS-CAM is not symmetric cell by cell, that is the reason of the capacitance difference between SL and $\overline{SL}$. Every critical line exhibits a reduction of the load, and the search lines reduction is due to the transmission logic that is finally loaded by one transistor. The most obvious reduction of the matchlie is caused by having less transistors connected to it.

4.2 Monte Carlo and Corner simulations

The Monte Carlo and corners simulations were made with $VDD = 1V$ and $27°C$ for a 3 σ variation. A Monte Carlo simulation of 1000 iterations was performed for each circuit to test the robustness and variability of the cell. The results observed on the Fig. 4.5 where the search time mean of NOR is $1581ps$ and for the proposed $924ps$. The variation is similar around the $185ps$, indicating a similar robustness of both cells.

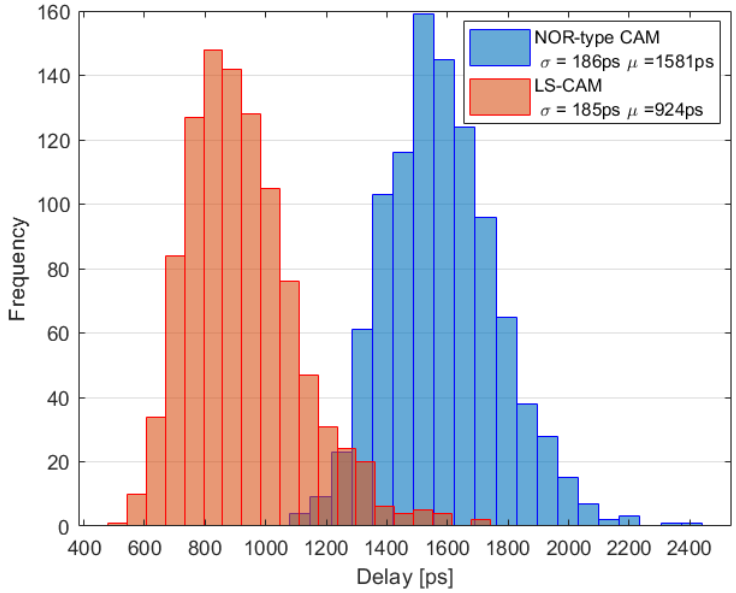


Figure 4.5: Monte Carlo delay results.

The LS-CAM offers a faster search operation than NOR-type CAM under process, voltage and temperature variations. The distribution of the Monte Carlo simulations around the corner simulations can be observed on Fig. 4.6. LS-CAM energy consumption is lower for all cases, and it is also more stable near to $1fJ$ for the most of the iterations. For the delay, it is noticeable that the sf and ss corners are near to the tt search times of the NOR-type CAM. The graph shows that slow nmos strongly decreases the performance of the circuit.

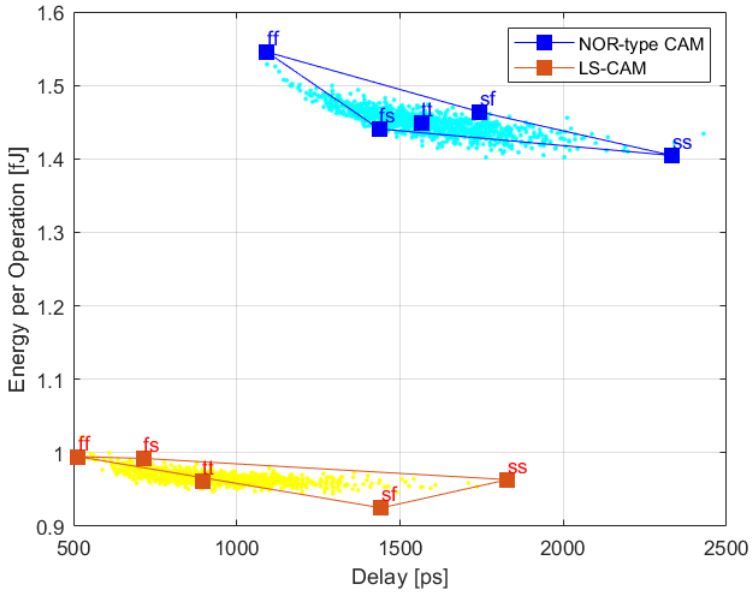


Figure 4.6: Energy vs Delay for Monte Carlo and corners simulations.

4.3 Voltage variations

The voltage variations were performed from 1.2V to the minimum value where the CAM works properly. The NOR-type CAM can operate from 0.3 V, while the LS-CAM fails on this voltage value and works from 0.4V. For lower values, $V_{DD} - V_T$ ML do not work correctly. In Fig. 4.7, it can be appreciated that the proposed CAM has a better performance from 1.2V to 0.6V. For lower values, the NOR CAM achieves better results. To be more specific, the NOR-type CAM has higher search time from 0.7V to 1.2V and consumes more energy from 0.5V to 1.2V.

The search time for the target voltage, 1 V, is 1565ps for the NOR-type CAM and 895.6ps for the LS-CAM. The energy consumption per operation of each bit-cell is 1.438fJ for the NOR-type CAM and 0.952fJ for the LS-CAM. This means that the NOR is 42% slower and consumes 33.8% more energy. From 0.7V to 1.2V, the LS-CAM is between 12.5% to 48.6% faster than NOR-type CAM. While for energy, from 0.6V to 1.2V the LS-CAM consumes from 17.9% to 36.6% less energy than conventional CAM.

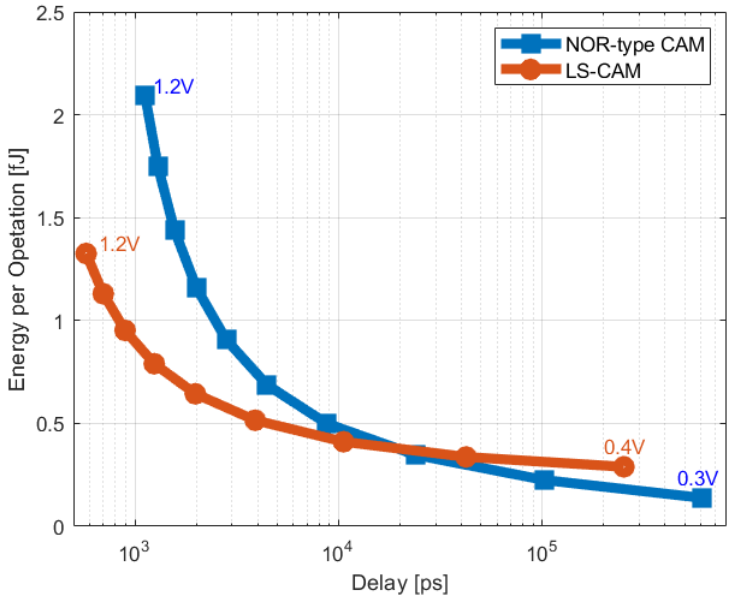


Figure 4.7: Energy vs Delay for voltage variations in steps of $100mV$.

4.4 Temperature variations

Temperature simulations were made at $V_{DD} = 1V$, in a range from $-20°C$ to $100°C$. As it is expected, the energy consumption increases with temperature. This trend is observable for both circuits in Fig. 4.8 where the graph shows a linear dependence. Nevertheless, the search time presents a different trend for the circuits. Fig. 4.9 shows a reduction on the search speed for the NOR-type CAM while there is an increase in the temperature. The LS-CAM exhibit a decreasing search time for higher temperatures.

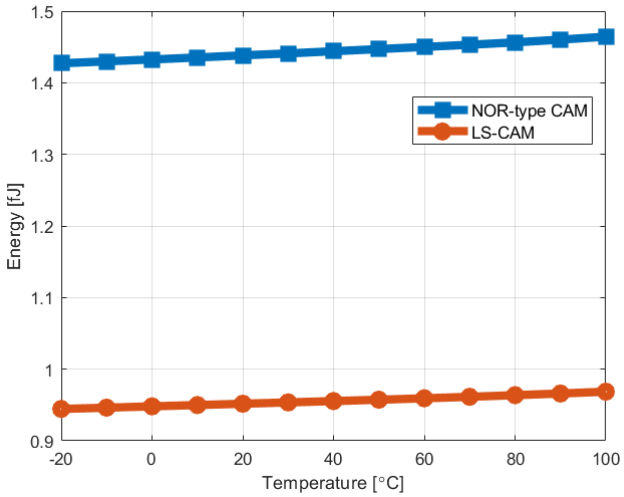


Figure 4.8: Energy vs Temperature.

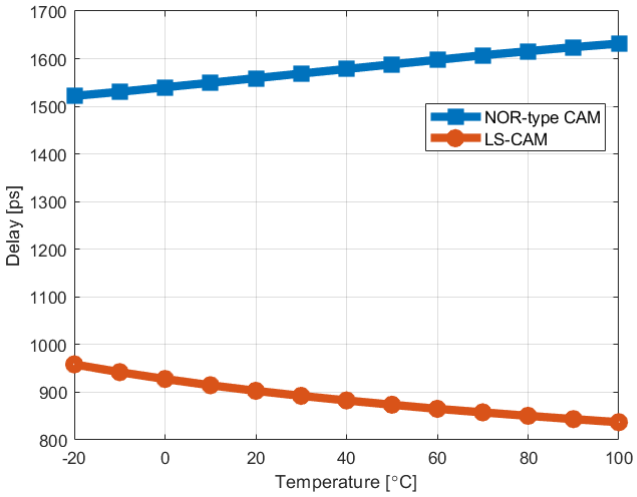


Figure 4.9: Delay vs Temperature.

These different trends can be caused by the low-swing approach. For a better analysis of the impact of the temperature, the evaluation and precharge time are plotted in Fig. 4.10. The time that the ML takes to precharge is not affected by the low-swing approach. This is due to the simulation conditions. To measure this time, the worst case scenario conditions were set. For the NOR-type CAM, the ML was set with an initial condition of $ML = 0$ and precharge of $ML = V_{DD}$. The LS-CAM was set to $ML = V_{DD}$, where it did not matter if it does not reach that value, it is just to simulate the worst case scenario. Therefore, the precharge time is not affected by the low-swing approach and increases smoothly with the temperature.

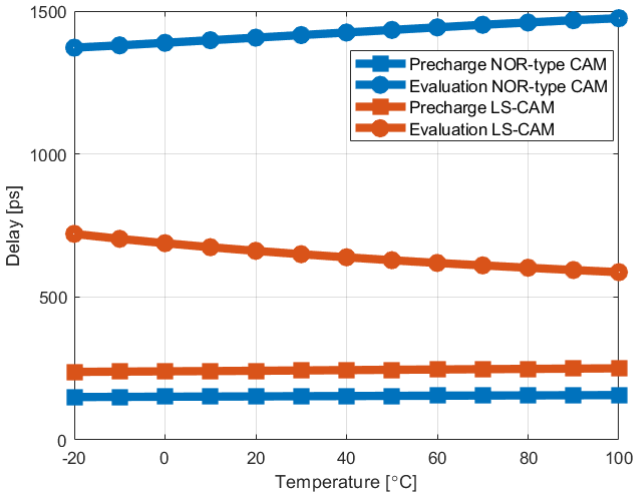


Figure 4.10: Precharge and evaluation times vs Temperature.

On the contrary, the evaluation time decreases with the temperature for the LS-CAM. This time it is influenced by the low-swing approach, since for the delay measurement, the LS-CAM should go from $0V$ to $V_{DD} - V_T$ and the temperature decreases the threshold voltage of the nmos [21]. Fig. 4.11 shows how at higher temperatures the ML reaches a higher value in the same time than the colder one. This triggers the inverter faster and gets a lower evaluation delay for higher temperatures.

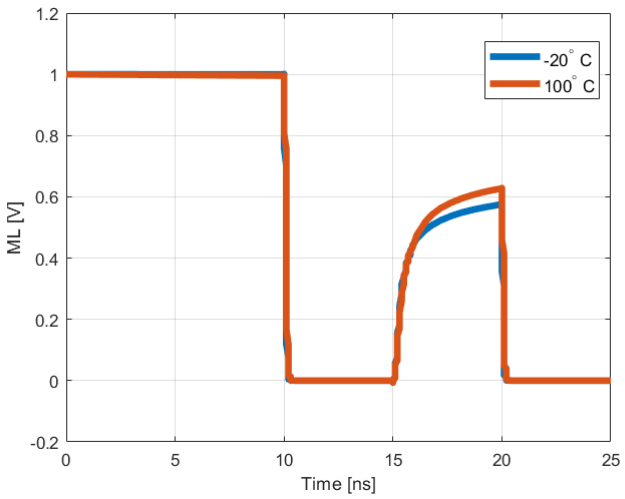


Figure 4.11: *Matchline* response for two temperatures.

4.5 Segments variations

Finally both circuits were tested for different sizes of segments, with the objective of knowing how the LS-CAM behaves in different number of segments. It also gives an idea of how suitable the circuit is for low energy techniques that require the segmentation of the word.

The structure of the word that was used for all the tests were 2 segments of 64 bits each. To maintain the size of the memory, the word length and the loads setting into the SLs and $\overline{SL}$s are the same. The other two configurations are a 4-segment and 8-segment word. To compute the match or mismatch of the word, an AND tree logic was used, and the same for a two segment word where the inputs of the logic are the MLs of the NOR-type CAM, and the inverter's outputs of the LS-CAM.

The measurement of the delay is similar to the 2-segment word. Just one segment performs a 1-bit mismatch and the search time is measured in the output of the AND tree. For the energy, one segment is set to perform a mismatch with the 50% of the bits. Fig. 4.12 and 4.13 shows the delay and energy results for the corners.

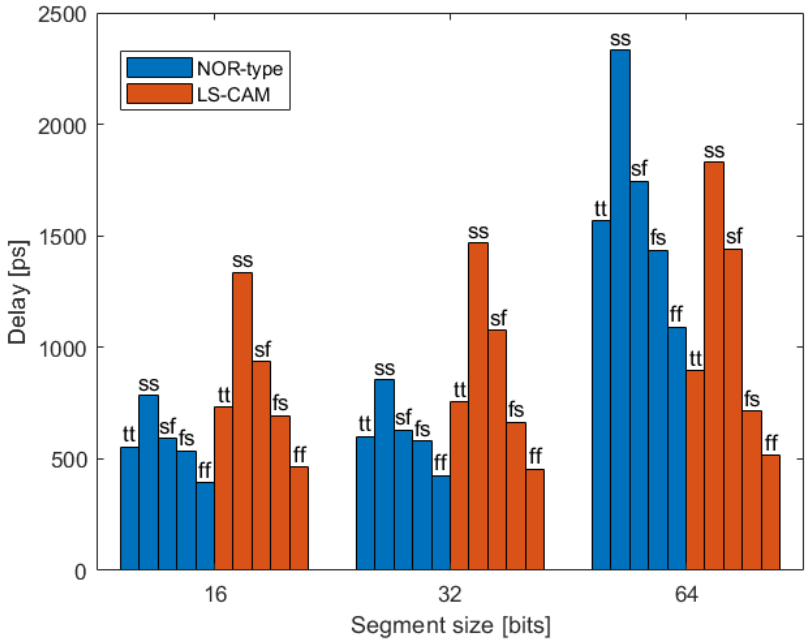


Figure 4.12: Search time for different segment sizes.

The delay graph shows that for smaller segments the LS-CAM has a worse performance than NOR-type CAM. On the typical simulation, for the 16-bit segment, the NOR-type is 180*ps* faster and for the 32-bit segment 154*ps*. Having a lower capacitance load benefits the full swing ML. In the case of the energy, the LS-CAM consumes less energy than NOR-type CAM for every configuration.

4.6 Comparison

Table 4.1 shows a performance comparison between the conventional and proposed cell with other works. The LS-CAM exhibits good results in the area compared to other CAM in 65nm. Also, it achieves an acceptable energy consumption, especially taking into account that it is similar to the 45 nm technology CAMs. Additionally, it has a competitive search delay, and the

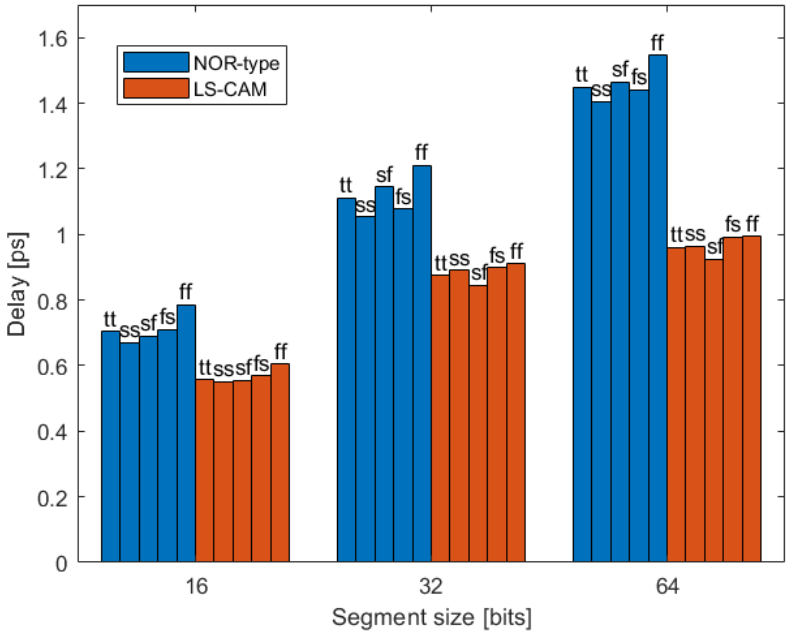


Figure 4.13: Energy per operation for different segment sizes.

[22] reaches impressive times but for smaller memories. Finally, the ternary CAM of [22] has a better performance, which achieves better search time and energy consumption, but the search time is improved by its higher supply voltage and its smaller capacity. If the LS-CAM employ a voltage supply of $1.1V$ the search time is reduced to $0.69ns$ which is almost the same of the TCAM. Also, with the size reduction, the energy to load the SLs will decrease and with it the energy per bit.

	Conventional	LS-CAM	[23]	[24]	[25]	[22]	[22]	[22]
Type	BCAM	BCAM	BCAM	TCAM	CAM	N-CAM	P-CAM	TG-CAM
Year	-	2024	2017	2019	2014	2019	2019	2019
Technology	65nm	65nm	45nm	65nm	65nm	45nm	45nm	45nm
Capacity	256x128	256x128	128x32	128X64	128x128	32x32	32x32	32x32
Supply [V]	1	1	1	1.1	1.2	1	1	1
Search delay [ns]	1.565	0.895	1.25	0.68	1.07	0.603	0.115	0.119
Energy/bit/search [fJ]	1.538	0.952	2.1	0.58	0.77	0.78	0.86	1
Cell area [μm^2]	3.42	3.44	-	-	3.8	-	-	-

Table 4.1: Performance comparison with other works.

Chapter 5

Conclusions

The designed circuit is an SRAM-based CAM cell circuit of 11 transistors. 6 of them are the SRAM normal transistors, 4 of them perform an XOR logic for the comparison and the last one connects the matchline to the power line to perform the match or mismatch logic. The design was made for $V_{DD} = 1V$ aiming for energy saving, while keeping an acceptable searching time and area.

The proposed and conventional circuits were made in 65 nm node technology. The layout of the LS-CAM was made to get a pattern each two cells, matching both by the match line and maintaining the area. LS-CAM cell only occupies 0.63% more area than NOR-type CAM. The LS-CAM precharge transistor is 3 times smaller but it requires a LO-skew inverter for each ML. So the area is almost the same in both cells.

The robustness of the cells is almost the same, but the LS-CAM achieves an $941ps$ average search time for 1000 Monte Carlo iterations while the conventional CAM reach a $1581ps$ average search time. The corners simulations shows that LS-CAM strongly depends on the nmos performance. However the objective is accomplished for every simulation, snd the energy consumption is lower for the proposed circuit.

The low-swing approach was properly fit into the circuit. However the operational voltage range is decreased due to the low-swing approach used which depends on the V_{DD} and the V_T of the transistor instead of a reference voltage circuit. The NOR-type CAM works from $0.3V$ and LS-CAM from $0.4V$. The LS-CAM has a longer search time from $0.4V$ to $0.6V$ but only the $0.4V$ has a bigger energy consumption. For the target voltage, the LS-CAM is

42% faster and consumes 33% less energy. The reduction of the voltage swing and the capacitance load was expected to offer more than half of the reduction of the power. However the results on energy per operation also takes into account the increment of the the switching frequency, and the leakage current between the SL and the power rails is bigger due to one nmos transistor instead of two transistors in stack.

In terms of temperature, the energy consumption keeps the same trend as the target voltage for all the temperatures. Contrary, the search time results shows an improvement of the LS-CAM performance for higher temperatures. The temperature decreases the threshold voltage of the nmos. Since the circuit strongly depends on nmos performance and low swing-approach, the circuit speed increases with the temperature.

Finally, the segment variations show that the LS-CAM reduces the performance for smaller segments. The energy results are better for the LS-CAM. This means that to save energy LS-CAM will be always better, but to keep the performance, larger segments are more suitable for the LS-CAM. Also the capacitance results show that the NOR-type adds more load to the search lines than LS-CAM. With this in mind and the fact that the operation steps and logic is the same, the pipelined schemes and hierarchical lines are suitable for both circuits. The LS-CAM performs better with bigger segments, so it will require less divisions and less extra circuitry to achieve the same energy saving. Thus, the LS-CAM offers lower energy consumption and good performance, specially for longer segment words.

Bibliography

[1] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (cam) circuits and architectures: a tutorial and survey," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, 2006.

[2] ——, "A low-power content-addressable memory (cam) using pipelined hierarchical search scheme," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1512–1519, 2004.

[3] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 4th ed. USA: Addison-Wesley Publishing Company, 2010.

[4] R. Sangireddy and A. Somani, "High-speed ip routing with binary decision diagrams based hardware address lookup engine," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 4, pp. 513–521, 2003.

[5] A. Annovi, L. Frontini, V. Liberali, and A. Stabile, "Design and characterization of new content addressable memory cells," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5.

[6] R. Karam, R. Puri, S. Ghosh, and S. Bhunia, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1311–1330, 2015.

[7] K. Pagiamtzis and A. Sheikholeslami, "Pipelined match-lines and hierarchical search-lines for low-power content-addressable memories," in *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference, 2003.*, 2003, pp. 383–386.

[8] S. Baeg, "Low-power ternary content-addressable memory design using a segmented match line," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 6, pp. 1485–1494, 2008.

[9] I. Arsovski and R. Wistort, "Self-referenced sense amplifier for across-chip-variation immune sensing in high-performance content-addressable memories," in *IEEE Custom Integrated Circuits Conference 2006*, 2006, pp. 453–456.

[10] H. Kim, M. Cho, S. Lee, H. S. Kwon, W. Y. Choi, and Y. Kim, "Content-addressable memory system using a nanoelectromechanical memory switch," *Electronics*, vol. 11, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/3/481

[11] Q. Dong, S. Jeloka, M. Saligane, Y. Kim, M. Kawaminami, A. Harada, S. Miyoshi, M. Yasuda, D. Blaauw, and D. Sylvester, "A 4 + 2t sram for searching and in-memory computing with 0.3-v v_{ddmin}," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 1006–1015, 2018.

[12] *Content Addressable Memory (CAM) Applications for ispXPLD Devices*, Lattice Semiconductor Corporation, July 2002, application Note AN8071.

[13] A.-T. Do, S. Chen, Z.-H. Kong, and K. S. Yeo, "A high speed low power cam with a parity bit and power-gated ml sensing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 1, pp. 151–156, 2013.

[14] A. T. Do, C. Yin, K. Velayudhan, Z. C. Lee, K. S. Yeo, and T. T.-H. Kim, "0.77 fj/bit/search content addressable memory using small match line swing and automated background checking scheme for variation tolerance," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 7, pp. 1487–1498, 2014.

[15] S.-U. Kim and K.-W. Kwon, "Charge recycling using folded match line in nor-type cam for reduced dynamic power," in *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2021, pp. 1–4.

[16] Z. Guo, D. Zhang, K. Zhang, M. Song, Y. Zhang, and L. Zeng, "A novel search-based compute-in-memory minimum values generation scheme for low-complexity ldpc min-sum decoding," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 7, pp. 3498–3502, 2024.

[17] S. Ahn and K.-w. Kwon, "Local nor and global nand match-line architecture for high performance cam," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 707–710.

[18] B.-D. Yang, Y.-K. Lee, S.-W. Sung, J.-J. Min, J.-M. Oh, and H.-J. Kang, "A low power content addressable memory using low swing search lines," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 12, pp. 2849–2858, 2011.

[19] ——, "A low power content addressable memory using low swing search lines," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 12, pp. 2849–2858, 2011.

[20] H. Miyatake, M. Tanaka, and Y. Mori, "A design for high-speed low-power cmos fully parallel content-addressable memory macros," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 6, pp. 956–968, 2001.

[21] J. Tzou, C. Yao, R. Cheung, and H. Chan, "The temperature dependence of threshold voltages in submicrometer cmos," *IEEE Electron Device Letters*, vol. 6, no. 5, pp. 250–252, 1985.

[22] T. V. Mahendra, S. Wasmir Hussain, S. Mishra, and A. Dandapat, "Performance analysis of n-cam, p-cam and tg-cam using 45-nm technology," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 621–625.

[23] T. Venkata Mahendra, S. Mishra, and A. Dandapat, "Self-controlled high-performance precharge-free content-addressable memory," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 8, pp. 2388–2392, 2017.

[24] K. Lee, G. Ko, and J. Park, "Low cost ternary content addressable memory based on early termination precharge scheme," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2019, pp. 1–4.

[25] A. T. Do, C. Yin, K. Velayudhan, Z. C. Lee, K. S. Yeo, and T. T.-H. Kim, "0.77 fj/bit/search content addressable memory using small match line swing and automated background checking scheme for variation tolerance," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 7, pp. 1487–1498, 2014.