An energy efficient content-addressable memory

Using a pipelined hierarchical search scheme to reduce power consumption

Fabian Finkeldey
ZITI, Institute of Computer Engineering
University of Heidelberg
Heidelberg, Germany
Finkeldey@stud.uni-heidelberg.de

***Abstract*—This report presents three techniques to reduce the power consumption in content-addressable memories (CAMs). The first approach is to break up the high capacitance match-lines into several segments and to pipeline the search-operation. Under the assumption, that most stored data-words won't match the search-word in their first segment, the search-operation can be discontinued for the most subsequent segments. This will reduce the power-consumption of search-operations significantly. The second approach is to introduce a hierarchical search-line scheme with different voltage-swing: Global search-lines run across the entire array with a low-swing signal and local-search-lines with lower capacitance will carry an amplified full-swing signal. Since the local-search-line-amplifiers will only be enabled, if the match-line-segment is enabled, only a few local-search-lines will be driven. This reduces the power consumption further. The third approach is to use a precharge-low-scheme for the matchlines. Hence the assumption, that most data-words won't match the search-word, this is more efficient than the conventional way of precharging all match-lines high. The application of these techniques will be shown in simulation and implementation in 180nm CMOS. The presented CAM has a dimension of 1024 x 144 bit, will reach a 7ns cycle time and will consume 60% less power, than a conventional nonpipelined and nonhierarchical architecture.**

Keywords—Assosiative Memory, content-addressable memory (CAM), hardware lookup, hierarchical search-lines, low-power, pattern-matching, pipelined hierarchical search scheme, pipelined match-lines, precharge-low-scheme

I. INTRODUCTION

Content-addressable memories (CAMs) compare search data against a table of stored data-words. The return-values are the addresses of the matching data. This search operation is much faster than software-search-algorithms, so it is widely used in search-intensive applications with tight timing-constraints, such as: caching in virtual address translation (translation lookaside buffer, TLB), address lookup in Internet routers and database acceleration.

Despite the wide range of applications, there are several reasons to keep the power consumption in mind, while designing a modern chip: In general, there are environmental benefits of reducing power consumption as well as economical advantages. For mobile applications it is key, to save as much battery life as possible, while high performance systems will benefit from a lower power draw by reducing the effort, space and cost associated with cooling-solutions.

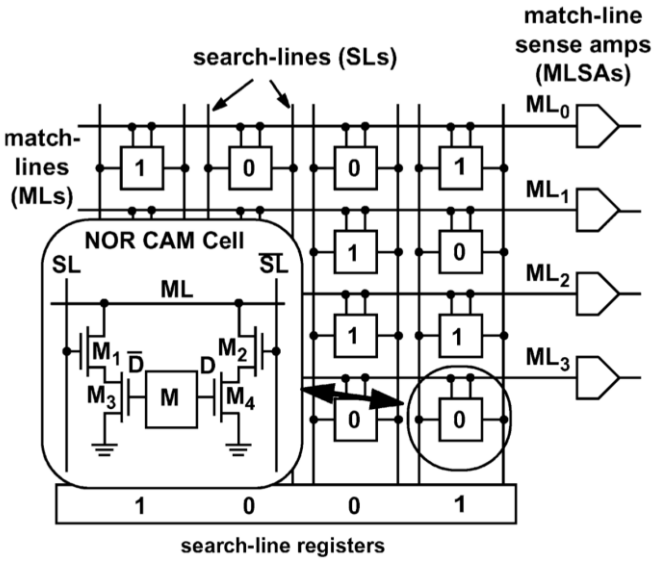

Figure 1: Simplified conventional CAM architecture with dimension of 4 x 4

Due to the highly parallel nature of a CAM search operation, which uses large amounts of circuitry in every cycle, it is especially power intensive. Since the required size of CAMs is increasing and the power consumption of conventional designs is scaling linear with the size, the power consumption of CAMs is rising.

When looking at a CAMs power consumption, there are two main components to consider: match-line and search-line power consumption. Both are highly active busses with a high capacitance running across the array. This document will propose to reduce match-line power consumption by introducing a pipelined architecture and a precharge-low-scheme. To reduce the search-line power consumption a hierarchical scheme is presented. By using these techniques, it is possible, to reduce the power consumption by 60% compared to a nonpipelined nonhierarchical design.

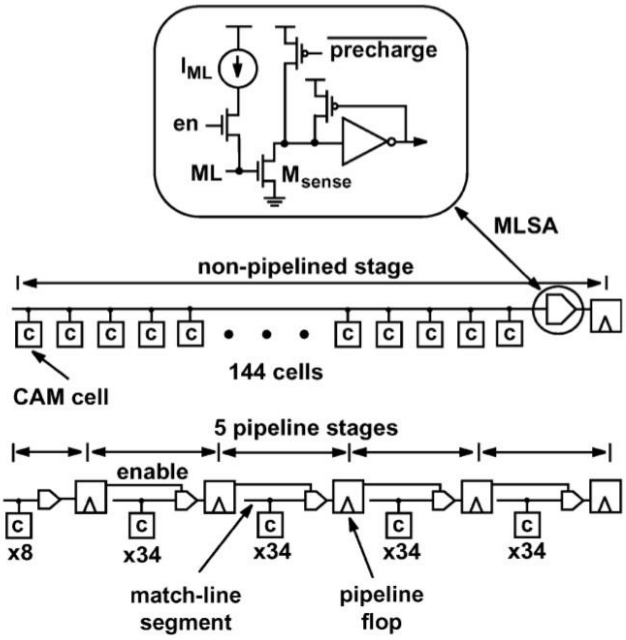

Figure 2: Match-line-sense-amplifier (MLSA, top) schematic, non-pipelined stage (middle) and pipelined architecture with 5 stages (bottom)

Section II of this document will give a short overview over the conventional design of CAMs. In Section III the pipelined match-line and the precharge-low-scheme are introduced. Section IV will present the hierarchical search-line architecture. Section V will show simulation data and measurements of a fabricated test chip to demonstrate the effectiveness of the proposed techniques. Finally, Section VI will discuss the trade-offs of the proposed architecture and Section VII will conclude this document.

II. Conventional CAM Design

Figure 1 shows a simplified block diagram of a conventional CAM design with four words of four bit length. The cells are implemented as NOR CAM cells and compare the data stored in the standard six-transistor SRAM cells *M* with the value assigned to the search-lines *SL*. In the figure the search-word is 1001, which matches the first line of data in this example. ML_0 will indicate that match by being on a high level at the end of the cycle.

To achieve this behavior, the cycle of the search operation (shown in Figure 3 a) starts with precharging all search-lines (*SL* and $\overline{SL}$) to ground and all match-lines (*ML*) to V_{DD}. The precharge of the search-lines will prevent a direct path current from the ML-precharge to ground, since the transistors M_1 and M_2 are switched off. The precharge of the *ML* will bring them all temporarily in match state.

After this precharge phase the search-word is driven onto the differential search-lines. This starts the comparison between the stored bit *D* and the corresponding bit of the search-word, which is driven to the search-line *SL*. In case of a mismatch between *D* and *SL*, there will be a pulldown-paths either over the transistors M_1, M_3 or M_2, M_4 from the match-line *ML* to ground. As a result,

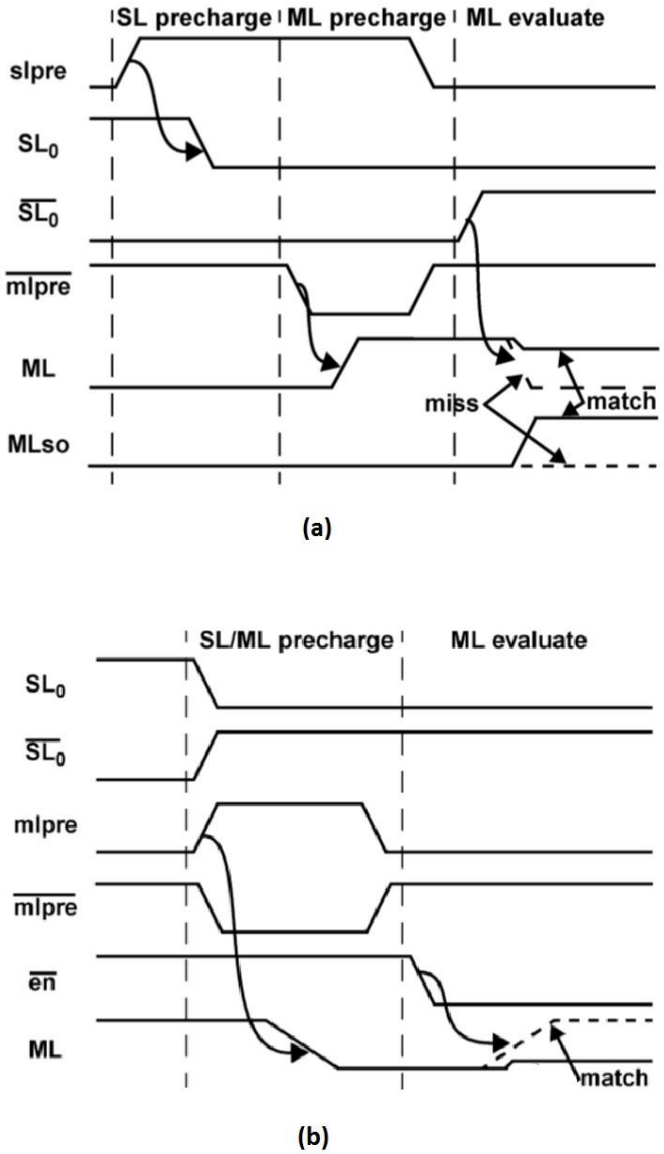

Figure 3: Timing diagram of conventional CAM (a) and CAM with precharge-low-scheme (b)

the entire match-line is pulled to ground, if there is a mismatch in any connected CAM cell. In case of a full match, where all bits of the stored data-word match the bits of the search-word, there won't be a path to ground in any of the connected CAM cells. So the match-line will stay in match-state.

To distinguish a match from a mismatch and to get a clean digital signal, there are match-line-sense-amplifiers at the end of every match-line. It is recommended to use the threshold voltage as the reference.

III. Reducing Match-Line Power Consumption

The match-lines are highly capacitive, as they are running horizontally across the entire array of CAM cells. Their total capacitance consists of the wire capacity and the diffusion capacity of the CAM cells. Under the assumption, that most data-words won't match the search-word, most match-lines will

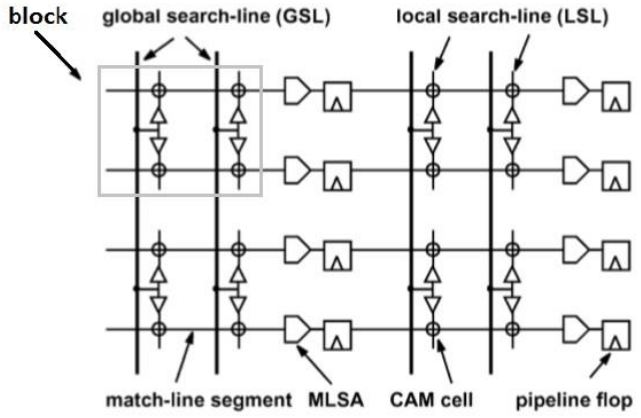

Figure 4: Pipelined CAM architecture with hierarchical search-lines

be precharged and discharged in every cycle, as described in section II. So they are one of the two main components of power consumption in the conventional CAM.

To reduce the power consumption, this section will propose a pipelined match-line architecture: As shown in Figure 2 the match-line can be broken up into segments. In this example, the 144 bit match-line is divided into 5 segments, with a flip-flop at the end of every segment. Only if a data-word matches in a segment, the subsequent segment is activated. So the entire data-word is compared to the search-word sequentially. Assuming, that most data-words will mismatch within the first segments, this leads to a significant reduction in power consumption, despite the overhead introduced by clocking additional flip-flops.

To reduce the power consumption of the match-lines further, the precharge-scheme can be modified. As described in section II, a precharge to V_{DD} leads to a high power consumption, if most data-words mismatch. In Figure 3 (b) a precharge-low-scheme is shown. In the precharge phase the matchlines are pulled to ground. This eliminates also the need to precharge the search-lines, since there is no risk of a direct path current to ground. So the search-lines can be driven differentially within the precharge-phase.

In the evaluation-phase the enable signal $\overline{en}$ enables a current source, which forces a constant current IML into the match-line. Match-lines, that are not in the match-state, will discharge this current to ground and so there is only a little increase of voltage due to the resistance of the path to ground. Match-lines in the match state won't have any path to ground and will collect charge. In this case the voltage increases to a point, where the match-line-sense-amplifier can detect a match. To ensure identical speeds on all segments, the current I_{ML} is devided in proportion to the segments length measured in the number of bits.

IV. Recuding Search-Line Power Consumption

Besides the match-lines, the search-lines are the second major contributor the power consumption of a CAM. As the match-lines, they are highly capacitive, due to their length and

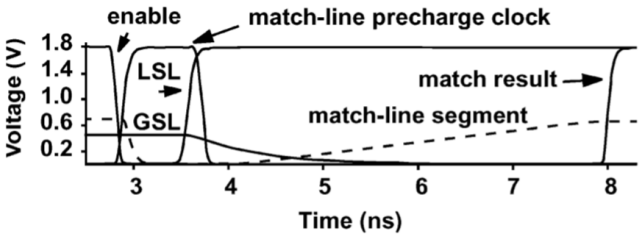

Figure 5: Signal waveforms for a full search cycle on a 34-bit segment.

the number of connected transistors. The overall capacitance of the search-lines consists of the wire capacitance and the gate capacitance of the connected transistors.

In a conventional design, the differential search-lines get precharged to ground, before driving them with the search-data. So one of the two differential lines has to be charged to V_{DD} in every cycle.

One approach to reduce the search-line power consumption is to drive them with a lower voltage than V_{DD}. This is an effective way of reducing the power consumption of the search-lines P_{SL}, since it depends on V_{SL} squared (1).

$$P_{SL} = C_{SL} * V_{SL}^2 \quad (1)$$

The disadvantage of this technique is, that the lower gate overdrive voltage reduces the speed of the CAM. To address this problem, this section proposes to introduce hierarchical search-lines in combination with the previously introduced pipelined match-lines (Figure 4): Global-search-lines, which will run across the entire array of CAM cells and are driven with a low-swing-signal and local-search-lines, which are using a low-swing-amplifier to amplify the signal of the global-search-lines to a full-swing-signal. The local-search-lines are shorter than the global-search-lines and so they are parting the CAM vertically.

Together with the technique of pipelining the match-lines, which leads to horizontal segmentation of the CAM cell array, the introduction of hierarchical search-lines creates several blocks. A block has the width of the match-line-segment and the height of the length of a local search-line.

The global-search-lines have to be driven for the entire CAM, since the search-data has to be broadcasted to all match-line-segments. The local-search-lines are only activated, if a match-line-segment is activated. Based on the assumption, that most match-line-segments are inactive, this results in a reduced power consumption, since the global-search-lines can operate at a low-voltage-swing and the amplifiers for the local-search-lines are rarely activated.

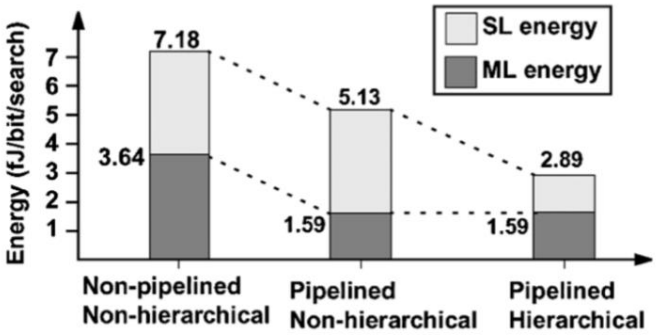

Figure 6: CAM search cycle power consumption for a nonpipelined, nonhierarchical design, a pipelined nonhierarchical design and a pipelined hierarchical design. The power consumptions are simulated for a 1024 x 144 bit CAM in a 1.8V 180nm CMOS process

V. Simulation

This section presents the simulation results of the proposed CAM architecture. The results include the effect of parasitics extracted from the layout. The timing is determined by a simulation of all sub circuits in HSPICE and power consumptions are determined by simulating the complete CAM at transistor-level in nanosim.

In determining the power consumption it was assumed, that there is only one match per search and the CAM is filled with uniformly distributed random data. The worst case power consumption would occur, when there is a large number of matches in a search cycle. Due to the rare occurrence of this case in most CAM applications it is not simulated or measured in this work. The simulation was done for a 1024 x 144 bit CAM in a 1.8V 180nm CMOS technology.

Figure 5 displays the simulated signal waveforms of a search-cycle in a 34-bit segment of the proposed CAM-design. The cycle begins by clocking the negative-edge-triggered low-swing-receiver to sense the current data on the global-search-line and by clocking the global-search-line flip-flop to sample the search-line-data for the next cycle. The receiver drives the local-search-line with a full-swing signal of the data sampled from the global-search-line. After this, the match-line sensing is activated and the match-line sensing clock activates the current-source that drives the current I_{ML} into the match-line-segment. In the case shown in Figure 5, the match-line is in match-state so the current I_{ML} leads to a rising voltage, since the match-line is able to collect charge. Only match-lines in match-state will reach a high enough voltage to trigger the match-line-sense-amplifier, before the current sources get turned off at the end of the cycle. This scheme achieves a result at the output of the match-line-sense-amplifier after a cycle time of 7ns.

Figure 6 shows a comparison of the power consumption between a conventional nonpipelined nonhierarchical architecture, a pipelined nonhierarchical architecture and a pipelined hierarchical architecture as proposed in this document. The conventional architecture consumes 7.18 fJ/bit/search, consisting of 3.64 fJ/bit/search for the match-line. The pipelined

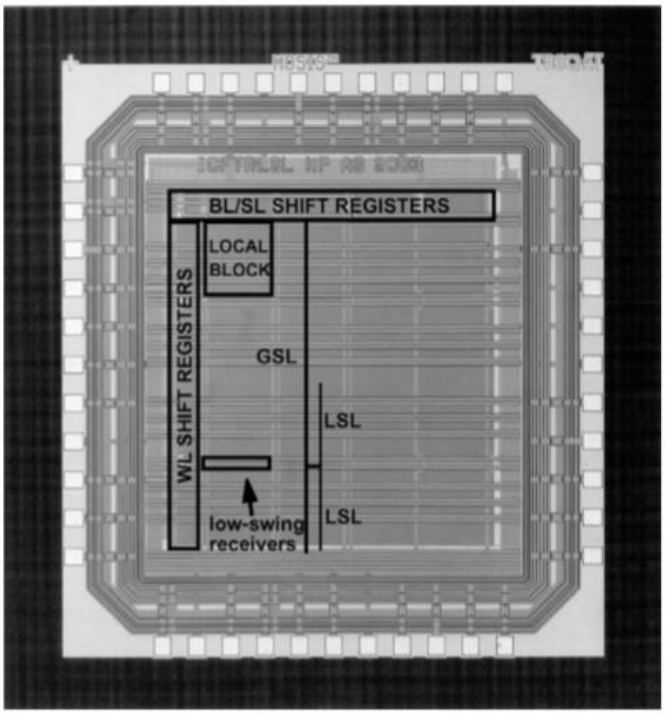

Figure 7: Die photomicrograph of the testchip implemented in a 1.8V 180nm CMOS technology

architecture reduces the match-line power consumption to 1.59 fJ/bit/search, while keeping the search-line power consumption unaltered. This leads to a reduced overall power consumption of 5.13 fJ/bit/search. Since most match-line-segments mismatch within the first 8-bit-segment, the following 136-bit-segments are rarely activated. So it could be expected, to see a drop in power consumption of around 96% (136bit/144bit). The simulated power consumption of the match-line is only reduced by 56%, which is due to the power consumption introduced by clocking the additional flip-flops at the end of every match-line-segment and the additional repeated circuitry for the match-line-sense-amplifier. In this implementation conventional master-slave flip-flops are used, which could be replaced with low-power flip-flops or pulsed latches to reduce the match-line power consumption further.

With the pipelined match-lines implemented, the search-line power consumption dominates the overall power-consumption, as shown in Figure 6. Adding hierarchical search-lines to this and using a voltage of V_{DDLOW} = 0.45V for the global-search-lines, the overall power consumption is reduced to 2.89 fJ/bit/search, consisting of a search-line-power consumption of 1.3 fJ/bit/search, which equates to a 63% reduction of search-line power consumption. Compared to a nonpipelined nonhierarchical architecture, this pipelined hierarchical architecture consumes 60% less power and reduces the power consumption from 7.18 fJ/bit/search to 2.89 fJ/bit/search.

Figure 7 shows a photomicrograph of a testchip implemented in a 1.8V 180nm CMOS technology. The testchip realizes a 256 x 144 bit CAM, which is smaller than the simulated one, due to limited silicon area. The implementation does not resort to any special devices.

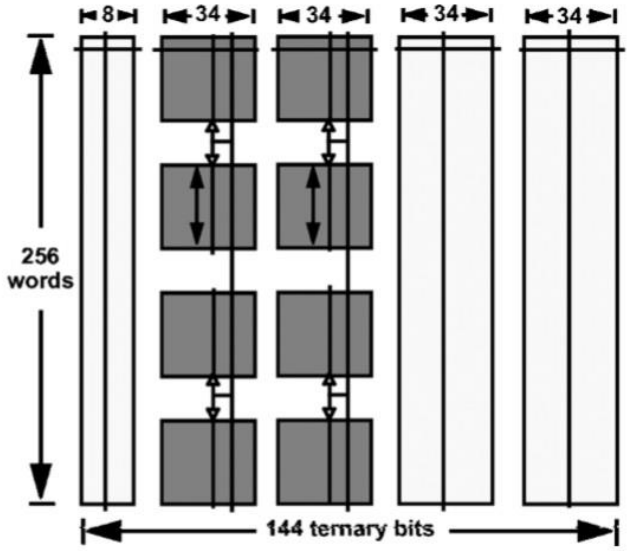

Figure 8: Testchip architecture. The matchlines are divided into 5 segments: the first with a length of 8 bit and the subsequent segments with 144 34 bits. For a direct comparison of the power consumption, the second and third segment (in grey) implement hierarchical search-lines

Figure 8 displays the overall organization of the testchip: The 144 bit match-lines are segmented into 5 parts. The first segment (left-most) has 8 bit, the subsequent 4 segments each have 34 bit. The local-search-lines are divided into 64 entry local blocks. To allow a direct comparison between the power consumption of hierarchical versus nonhierarchical search-lines, the second and third segment implement a hierarchical architecture while the fourth and fifth use conventional nonhierarchical search-line-scheme.

Figure 9 shows the comparison between simulated and measured power consumption of the hierarchical search-lines depending on the number of activated blocks. The measured energy values are within 5% of the simulated values.

VI. DISCUSSION

A. *Further Optimization*

The pipelined hierarchical search scheme presented in this document does not depend on a specific implementation of the match-line-sense-amplifier or a specific match-line-sensing-scheme. This leaves room for further optimization. Furthermore, other techniques like bank selection [3],[4] or selective precharge [5] can be applied to reduce the power consumption.

B. *Benefits and Disadvantegs depending on the application*

The presented architecture can, as shown in previous sections, result in a significant reduction of power consumption. Depending on the design, especially on the size and number of the local-search-lines and the size and number of the match-line-segments the costs are an increased chip area, an additional

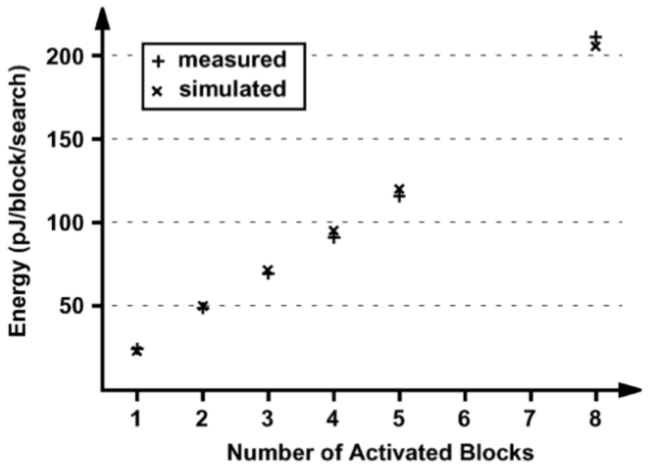

Figure 9: Measured power consumption of the hierarchical search-lines versus the number of activated blocks

latency introduced by the pipeline and a higher sensitivity to noise, due to the introduction of low-swing signals and additional amplifiers. While it is possible, to minimize the area-overhead to a rather small amount (it is around 6% in the presented design), the other disadvantages might disqualify the proposed architecture for a specific application.

The simulated and measured reduction of energy consumption of about 60% was under the assumption, that most data-words will mismatch the search-word within the first 8 bit and that the memory is filled with uniformly distributed data. While it is safe to assume this is true for most applications, there are other CAM applications where this might not be the case. An IPv4-lookup-table for example might deliver a high number of matches within the first 24 bits.

VII. CONCLUSION

This document presented three techniques to reduce power consumption of content-addressable-memory (CAM): The introduction of a pipelined match-line, the use of a precharge-low-scheme instead of the conventional precharge-high-scheme and the hierarchical search-line-scheme. The goal was to activate fewer parts of the chip than in a conventional architecture and thus save power.

With a simulation and measurements from a 180nm testchip it was shown, that these techniques can reduce the power consumption by 60%. These results were under the assumption, that most data-words won't match the search-word within the first segments.

Finally possibilities for further optimization and the benefits and disadvantages of the presented design were discussed. Depending on the application, the energy savings might be lower, due to a specific data-distribution. In other cases, the proposed architecture might be disqualified by the slight area-overhead or the increased noise-sensitivity.

REFERENCES

[1] K. Pagiamtzis, A. Sheikholeslami, “A low-power content-addressable memory (CAM) using pipelined hierarchical search-scheme”, IEEE Journal of solid-state Circuits, 2004

[2] K. Pagiamtzis, A. Sheikholeslami, “Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey”, IEEE Journal of solid-state Circuits, 2006

[3] F. Zane, G. Narlikar, and A. Basu, “CoolCAMs: Power-efficient TCAM’s for forwarding engines,” in Proc. IEEE INFOCOM, vol. 1, 2003, pp. 42–52.

[4] G.Kasai,Y.Takarabe,K.Furumi,andM.Yoneda,“200MHz/200MSPS 3.2 W at 1.5 V Vdd, 9.4 Mbits ternary CAM with new charge injection matchdetectcircuitsandbankselectionscheme,”inProc.IEEECustom Integrated Circuits Conf., 2003, pp. 387–390

[5] C. A. Zukowski and S.-Y. Wang, “Use of selective precharge for lowpowercontent-addressablememories,”inProc.IEEEInt.Symp.Circuits and Systems, vol. 3, 1997, pp. 1788–1791.