Highly Parallel and Energy-Efficient Exhaustive Minimum Distance Search Engine Using Hybrid Digital/Analog Circuit Techniques

Chang Ki Kwon and Kwyro Lee

Abstract—A minimum distance search engine (MDSE) is presented as a hardware accelerator for various exhaustive pattern-matching systems. This chip executes highly parallel computations of $L_1$-norms between an input query and stored multiple reference records, and searches for the minimum distance among them in a highly parallel fashion. Our architectural-level estimation shows that this MDSE can reduce energy dissipation by orders of magnitude as the number of records increases, compared with the conventional systems. We have designed a prototype 4-bit 8-word MDSE composed of merged memory logic (MML) and digital/analog-mixed winner-take-all circuit (DAM-WTAC) by using hybrid digital/analog circuit techniques. It was fabricated with a 0.6-$\mu$m single-poly triple-metal CMOS technology. Experimental results show that our chip works properly at 3 V/10 MHz and has approximately four times larger throughput as well as four times higher energy efficiency, compared with the existing 8-bit microcontrollers.

Index Terms—Digital/analog-mixed circuit, low-power design, memory, static CMOS combinational circuit.

I. INTRODUCTION

Traditional von Neumann computers are suited for computational tasks that can be described by a sequential deterministic algorithm. Although the progress in VLSI technology has substantially improved the performance of single-processor systems, these are still not fast enough to perform certain applications within a reasonable time period, such as real-time control optimization, real-time speech recognition, real-time pattern matching, and so forth [1]. Moreover, due to the separation of a processor and memory units, the heavy traffic between them is apt to dissipate quite a large power. As a matter of fact, 50–80% of total power dissipation is originated from the memory accesses both for custom-hardware [2] and general purpose processors [3]. This has definitely become one of the limiting factors in enhancing performances of portable systems.

To increase the throughput, several parallel computers have been proposed [1]. However, as long as main memories are external to the processors, the parallel computers can not significantly save the total energy necessary to fulfill a given task solely by parallelism, even with architecture-driven supply voltage scaling [4].

For further improvement in both throughput and energy dissipation, recent VLSI technology focuses on merging memories and high-density logic in a chip [5]. This so-called Merged Memory Logic (MML) technique is opening new horizons even for general-purpose computer systems in terms of speed and energy performance. In addition, greater performance gains are expected, if we adopt special hardware accelerators for reducing unnecessary memory accesses when the task is a simply repeated data-sequence of regular algorithms. For example, a number of scalar/vector quantization encoding engines has been developed by using analog [6], digital [7], [8], and hybrid digital/analog [9] techniques. The analog implementation showed superior energy efficiency but has the limited accuracy of analog computation. The digital implementations demonstrated highly parallel operation with vast multidimensional data but have low throughput due to the bit-serial word-parallel search. In our previous work [9], we have suggested an energy-efficient exhaustive minimum distance search engine (MDSE), which can be easily extended for scalar/vector search and has relatively high throughput due to the bit-parallel word-parallel search.

In this paper, we present our implementation and experimental results of the prototype MDSE chip. This chip computes $L_1$-norms between an input scalar-query and stored multiple scalar-records by using digital MML, and searches for the minimum distance among them by using digital/analog-mixed winner-take-all circuit (DAM-WTAC), both in highly parallel manners. Therefore, for adapting our chip to the different applications, proper VLSI architectures using the proposed circuits should be more considered like other implementations [6]–[8].

II. IMPLEMENTATION OF MDSE

In general, MDSE is organized with three elements: local memories to store input queries and reference records, distance calculators, and minimum distance finders. Depending on the integration level of these elements in a single chip, the MDSE architectures can be classified into three generic types, as shown in Fig. 1.

Let us assume that $M$ input queries and $N$ reference records are initially stored in the external main memories. In type A-MDSE shown in Fig. 1(a), $M$ queries and their respective $N$ reference records are sequentially accessed from the main memories and processed in the data-path processor. However, by prefetching the reference records into the local memory in type B-MDSE shown in Fig. 1(b), significant energy saving can be achieved. Note, however, that throughput is not improved in comparison with type A-MDSE. In type C-MDSE [6]–[9], much improvement both in throughput and energy can be obtained for a given query by calculating distances between $N$ records simultaneously, while this requires large local memories as well as many distributed processing elements. Notice that the throughput of type C-MDSE is about $N$ times larger than that of the others. Moreover, notice that the internal memories of type C-MDSE are accessed only once as long as the reference records remain unchanged. This results in the drastic energy reduction when both of $N$ and $M$ is large.

A. Architectural Level Estimation

For architectural level estimation in practical cases, we first assume that the supply voltage ($V_{dd}$) is fixed. In addition, it is assumed that the clock frequency ($f_{clk}$) is scaled linearly below the maximum frequency at which the external SRAM is accessed and that $f_{clk}$ can be gated according to the power-down strategies. Then, the maximum clock frequency is only needed to compute the minimal number of parallel chip sets.

All the power and area libraries have uniformly been adapted for a 0.6-$\mu$m CMOS technology, operating at 5 V. For the SRAM, we have used the power model of Fujitsu’s 1-Mbit-SRAM [11] whose maximum frequency is 100 MHz at 5 V. The other power models have been adopted from the developed models of embedded RAM, registers [12] and data paths [13]. They have then been slightly modified by constant

Manuscript received July 10, 1999; revised March 20, 2000. This work was supported in part by the Korea Science and Engineering Foundation through the MICROs Center, and in part by IDEC and Samsung Electronics Company. The authors are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Taejon 305-701, Korea.

Publisher Item Identifier S 1063-8210(00)09467-1.
Fig. 1. Block diagrams of minimum distance search engine (MDSE). (a) Type A—MDSE. (b) Type B—MDSE [10]. (c) Type C—MDSE [6]–[9].

E-field scaling theory [14]. For energy estimation, we have multiplied the average dynamic power found in [9] by the time period \(=1/f_s\) required to find the minimum distance per one query. For area estimation, we have used a 0.8-\(\mu\)m COMPASS library [15], which has been also scaled similarly to the given technology. Finally, we assume that 10 MHz \(=f_s\) sampled queries are compared in brute-force manner with \(N\) records of 4-bit unsigned integer representation.

Based on the aforementioned assumptions, the estimated chip-area and the energy dissipation per query are calculated, as shown in Fig. 2(a) and (b), respectively. In Fig. 2(a), notice that type C-MDSE does not waste the memory area while the others require many chip sets including unnecessary memories to meet the required throughput within the maximum access frequency. Finally, we assume that 10 MHz \(=f_s\) sampled queries are compared in brute-force manner with \(N\) records of 4-bit unsigned integer representation.

Based on the aforementioned assumptions, the estimated chip-area and the energy dissipation per query are calculated, as shown in Fig. 2(a) and (b), respectively. In Fig. 2(a), notice that type C-MDSE does not waste the memory area while the others require many chip sets including unnecessary memories to meet the required throughput within the maximum access frequency. Finally, we assume that 10 MHz \(=f_s\) sampled queries are compared in brute-force manner with \(N\) records of 4-bit unsigned integer representation.

Based on the aforementioned assumptions, the estimated chip-area and the energy dissipation per query are calculated, as shown in Fig. 2(a) and (b), respectively. In Fig. 2(a), notice that type C-MDSE does not waste the memory area while the others require many chip sets including unnecessary memories to meet the required throughput within the maximum access frequency. Finally, we assume that 10 MHz \(=f_s\) sampled queries are compared in brute-force manner with \(N\) records of 4-bit unsigned integer representation.

Based on the aforementioned assumptions, the estimated chip-area and the energy dissipation per query are calculated, as shown in Fig. 2(a) and (b), respectively. In Fig. 2(a), notice that type C-MDSE does not waste the memory area while the others require many chip sets including unnecessary memories to meet the required throughput within the maximum access frequency. Finally, we assume that 10 MHz \(=f_s\) sampled queries are compared in brute-force manner with \(N\) records of 4-bit unsigned integer representation.

Based on the aforementioned assumptions, the estimated chip-area and the energy dissipation per query are calculated, as shown in Fig. 2(a) and (b), respectively. In Fig. 2(a), notice that type C-MDSE does not waste the memory area while the others require many chip sets including unnecessary memories to meet the required throughput within the maximum access frequency. Finally, we assume that 10 MHz \(=f_s\) sampled queries are compared in brute-force manner with \(N\) records of 4-bit unsigned integer representation.

B. Merged Memory Logic Circuit

We have proposed a novel MML circuit [9] designed to calculate the absolute difference value in parallel between an \(n\)-bit input query \(=Y\) with \(n\)-bit records \(=X\) stored in SRAM. Fig. 3 presents a 1-bit MML with a 1-bit SRAM cell. Hence, the \(n\)-bit MML can be constructed by cascading \(n\) 1-bit MML like an \(n\)-bit ripple carry adder.

The operation of MML is as follows. When every CFG1*CSi becomes high, the input data \(Y^i\) and \(Y^i\) can be inserted. As RESET falls to zero, the circuits start to perform two subtractions simultaneously in the interconnect lines are not considered in Fig. 2(b). Taking these into account would result in an even larger energy reduction.

Fig. 2. Estimated results of total area and energy at \(V_{dd} = 5\ V, f_s = 10\ MHz,\) and \(f_{max} = 100\ MHz\). (a) The estimated chip area. (b) The estimated energy dissipation per query.

Fig. 3. The \(i\)th bit merged memory logic (MML) circuit diagram.
and \( X - Y \) and \( Y - X \). Using the most significant bit (\( \equiv C_n \)) of these two operations, the multiplexer selects the positive one. Therefore, the output (\( \equiv D \)) gives the absolute difference value (\( \equiv |X - Y| \)). The detailed operation and equations are described in our previous publication [9].

Our proposed MML has the following features: First, it is very compact, modular, and scalable. This is because the first half adder is a combination of the XOR with positive Manchester carry chains having transmission gate structures. Second, for low power consumption, it is designed by using signal feedback technique [16] and charge-recycling scheme [17]. Finally, our MML can be easily programmed to execute various functions such as subtraction, addition, and absolute difference calculation, via simple controls of the least significant bits and the most significant bit of carry signals.

When the MML is expanded similar to the configuration of conventional memories, the searching speed is limited by the slew-rate on the bit lines. Hence, for a faster exhaustive search, cross-coupled repeaters should be inserted at the bit-lines and turned on only during the searching period.

### C. Digital/Analog Mixed Winner-Take-All Circuit

A variety of circuits that perform the WTA function have been reported recently for analog and digital signal processing systems. Analog WTA circuits (AWTAC) that rely almost invariably on voltage followers (VF) are ubiquitous in virtue of small size and fast speed, especially for analog systems. Several modifications were proposed to enhance the performance limited by VF’s low resolution-gain [18]. However, most of their gain stages still suffer from considerable change due to the process, temperature, and supply voltage variations.

We have proposed a novel AWTAC [9], where the resolution gain is solely determined by the (\( W/E \)) ratio of transistor M1 to transistor MR1, both of which operate in the triode region. We have also extended this AWTAC to DAM-WTAC, as shown in Fig. 4, by adding a replica-bias circuit and a simple digital-to-analog input stage that converts the digital voltages to analog currents. The summed current is linearly converted into the analog voltage at the gate of the transistor MR1 by the negative feedback (MC1, MSF1, and MR1). After the WTA process, transistor MPF1 forces transistor MSF1 to be turned off, and stabilizes the output stage by the positive feedback (MPF1, MP1, and MP2) [6].

The proposed DAM-WTAC is basically classified as an \( O(N) \) system for low precision, but can be easily extended to logarithmic architecture of complexity \( O(N \log_2 N) \) for higher precision, or for a larger database at the expense of the unit-resolution reduction, as analyzed in [18]. Simultaneously, it is possible to reduce some delay by increasing the static currents through the output stages.

### III. EXPERIMENTAL RESULTS

We have implemented a prototype 4-bit 8-word MDSE composed of MML and DAM-WTAC with a 0.6-\( \mu \)m single-poly triple-metal CMOS process. Fig. 5 shows the microphotograph of the fabricated chip. The performance of the chip is summarized in Table I.

Fig. 6 compares the measured power dissipation per throughput or energy dissipation per query of our 4-bit 8-word MDSE with the 8-bit microcontroller data [10] at 3 V. Notice that our MDSE has about four times higher throughput and dissipates about four times less energy, because the referred throughput of the microcontrollers must be divided by eight for finding one minimum distance record. The performance gain will increase drastically as the record size becomes larger, as described in Section II-A. Furthermore, comparing our MDSE with the other parallel engines [6]–[8], we have observed that energy dissipation per total bit of our chip is comparable and that throughput per query of ours is relatively high due to the bit-parallel word-parallel search.

Fig. 7 shows the measured transient response of our DAM-WTAC in the worst case. Notice that the high pulse on \( V_{rst} \), resetting all the outputs to zero, marks the beginning of the WTA cycle. The measured transient response shows the worst-case delay below 60 ns.
Fig. 6. Energy dissipation per query of our MDSE compared with various 8-bit microcontroller data at 3 V.

Fig. 7. Measured transient response of our DAM-WTAC in the worst case.

IV. Conclusion

A highly parallel and energy-efficient exhaustive MDSE has been presented. The estimated energy dissipation in architectural level shows that our MDSE reduces energy dissipation by orders of magnitude, as compared with the conventional systems. This is because it avoids the repeated record-accesses when the number of records is quite large. To demonstrate the feasibility of our MDSE, a prototype 4-bit 8-word chip composed of MML and DAM-WTAC has been implemented with a 0.6-μm single-poly triple-metal CMOS process. The prototype chip works properly at 3 V/10 MHz and dissipates about 8 mW.

It is expected that both the chip and the hybrid digital/analog circuit techniques will be useful especially for portable multimedia applications such as database search, vector quantization, and pattern matching.

Acknowledgment

The author would like to thank MICROS, IDEC and Samsung Electronics Company for their support. They would also like to thank the reviewers for their valuable comments and Dr. K. Kim, Samsung Electronics Company, for useful discussion.

References