Extended Abstracts of the 2009 International Conference on Solid State Devices and Materials, Sendai, 2009, pp454-455

A Binary-Tree Hierarchical Multiple-Chip Architecture for Real-Time Large-Scale Learning Processor Systems

Yitao Ma and Tadashi Shibata

Department of Electrical Engineering and Information System, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Phone: +81-3-5841-6797 E-mail: mrmyt@if.t.u-tokyo.ac.jp, shibata@ee.t.u-tokyo.ac.jp

1. Introduction

Real-time learning function is becoming increasingly important in time-critical applications under complicated and variable situations, such as in automotive car control, robotic control, video surveillance and so forth. Because of the heavy computational cost arising form a large number of iteration cycles using a large volume of learning samples, the real-time performance is not feasible only by software approaches. Therefore various kinds of hardware/software hybrid implementation [1,2] and silicon intellectual property [3] have been investigated. A dedicated VLSI processor [4] has been developed for K-means learning algorithm, and a response time as short as milliseconds is demonstrated. However, due to the embedded memory architecture, the maximum data size is limited by the chip size and not applicable to massive data learning, for instance, as required in large-size image analysis.

In order to solve the problem, a multiple-chip architecture has been proposed [5], in which the data size can be scaled up by just increasing the total number of chips in the learning system without scarifying the speed performance. However, several critical problems have been revealed in our previous approach: (1) very complicated external controls of inter-chip data transmission; (2) the necessity of chip design change according to the scale up of the entire system; (3) a large overhead area in the chip that is necessary for each chip to adapt itself to a specific stage in the hierarchy.

In this paper, we have developed a new binary-tree hierarchical multiple-chip architecture to solve all these three problems, while retaining the merits of our previous work [5]. By employing a binary-tree chip connection structure, the real-time K-means learning system can be scaled up to any size using chips all having an identical design and minimal complexity. In addition, the inter-chip data transmission is automatically performed without any complex external control. Furthermore, it is also possible to extend the capability of the system for real-time recognition processing by just adding a winner-take-all circuitry to each component chip. A proof-of-concept chip for only the learning function part was designed in a 0.18-μm 5-metal CMOS technology and sent to fabrication. Its operation has been verified by NanoSim simulation.

2. K-means Learning Algorithm

The K-means clustering algorithm is a very powerful learning tool. In the algorithm, a large quantity of learning samples given in the form of multidimensional vectors are autonomously partitioned into limited number (K) of clusters according to their feature similarity. Each centoid represents the feature of the class and the set of centroid vectors are utilized as learning results. The procedure is illustrated in Fig. 1 for K=3 as a example. Initial seed vectors are chosen at the beginning. Then the distance from each sample vector to every seed vector is calculated to allocate all sample data into their nearest neighbor seed vectors, thus classifying the entire data points into K clusters. Then the centroid vector is calculated as the mean of each cluster, which serves as a new seed vector in the following clustering step. The process is repeated until all seed vectors become stable.

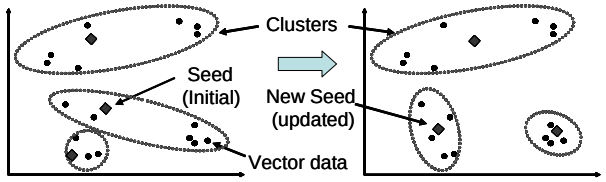

Fig. 1 K-means learning algorithm for K=3. Two steps in the beginning of iteration are shown.

3. Binary-Tree Hierarchical Multiple-Chip Architecture

Fig. 2 shows the system configuration of the K-means Binary-tree hierarchical multiple-chip architecture. This system can be unlimitedly extended by connecting VLSI chips having the same design configuration shown in Fig. 2 (a). A single chip consists of a set of internal memories for storing learning sample vectors, a cluster ID mask (CIMD) circuit for passing through learning samples having an identification number specified by the system, a local accumulator (LA) for interim result accumulation during centroid calculation, a global centroid unit (GCU) for generating mean vectors, variable delay circuits for managing inter-chip data transmission, fixed delay circuits for controlling pipeline operation, and massively parallel distance units (DUs).

In this architecture, learning sample vectors (64-dimension, 8b for one element) are stored in N cores of SRAM banks, and each core has an array of 64 x 32 SRAM cells. And each sample vector has its own cluster identification (CID) indicating the cluster it belongs to, which is stored in CIDM circuit and updated at the end of every iteration. During each iteration cycle, the learning sample vectors are element-serially downloaded from the SRAM to

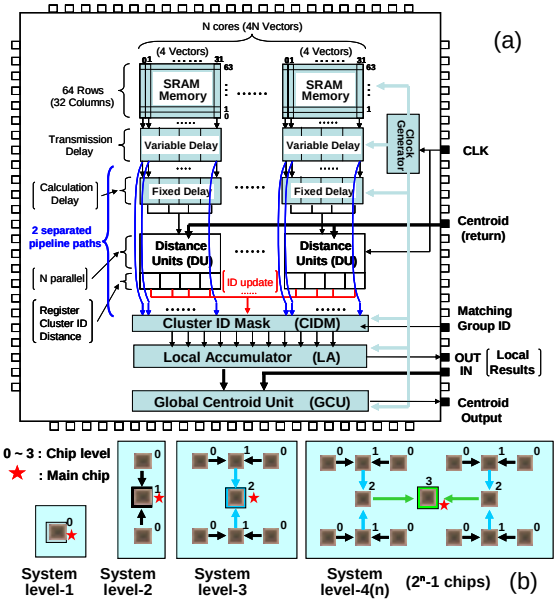

Fig. 2 System organization of K-means binary-tree hierarchical multiple-chip architecture: (a) on-chip organization, (b) binary-tree hierarchical multiple-chip organization.

DUs through two separate paths. One is the indirect path through CIDM, LA, and GCU for centroid calculation and the other is the direct path through fixed delay circuits waiting for the centroid results. In order to complete the centroid calculation of the vector data which belong to the same cluster but locate in different chips, interim results of every chip have to be brought together to the main chip to obtain the final centroid result. For reducing the number of necessary pins of the mainchip which becomes the bottle neck for inter-chip data transmission, a binary-tree hierarchical data transmission scheme has been employed. In this scheme, the interim results of each chip are accumulated and sent firstly from level 0 chips to level 1 chips, then to level 2 chip,……, and finally to the main chip. This will cost a few more clock cycles during the pipeline process for centroid calculation in each iteration, but the increase in the entire computational cost is negligible. The system scale can be doubled by increasing a single level in the system as shown in Fig. 2 (b). The timing of sending and receiving interim results to and from chips in different levels can be adjusted by the variable delay circuits.

For accelerating the system, DUs are driven with the system clock, while the centroid computation part of this architecture is driven with a four times lower clock frequency to absorb the inter-chip transmission delay. As a result, the distance calculations for all 4 sample vectors in one core are accomplished during one centroid computation period. Therefore, no idling operation of local chips happens and very efficient processing has been achieved. Because the distance can be calculated locally in each chip, the maximum data capacity that single chip can handle will be easily scaled up by increasing the number of cores employing more advanced technology nodes. In addition, it should be noted that, using the distance values (similarity)

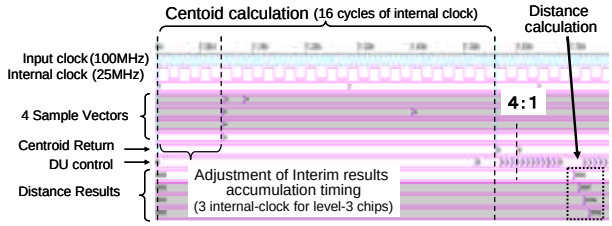

Fig. 3 NanoSim simulation results of single chip operation (chip-level = 3)

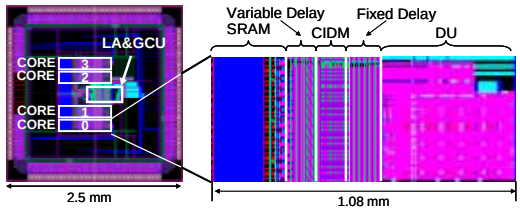

Fig. 4 Layout of the prototype chip with 4 cores for 16 learning sample vectors.

stored in DUs, the recognition function can be easily implemented by inserting a winner-take-all (minimum-selection) circuitry between DUs and CIDM.

4. Design and Simulation Results

The proof-of-concept chip for only learning function was designed in a 0.18-μm, 5-matal CMOS technology using cadence tools (icfb). The single chip operation at 100MHz is demonstrated by NanoSim simulation under a power-supply voltage of 1.8V (Fig. 3). The layout for four cores (16 learning vectors) circuitry produced by hand-layout is shown in Fig. 4. It takes 16 internal-clock cycles generated with four times longer period than the system clock to calculate the centoid vector. And the timing of interim results accumulation is adjusted by variable delay circuits (3 internal-clocks for level-3 chip). The results of distance calculation between one element of a returned centroid vector and the four elements of four different sample vectors which come through the delay circuit consecutively are overwritten to the register using system clock.

5. Conclusions

A binary-tree hierarchical multiple-chip architecture was proposed for real-time learning and recognition function of image processing. The NanoSim simulation results confirmed the operation of the proof-of-concept chip.

References

[1] A. Toriumi and Y. Hirayama, *Advanced Metallization and Interconnect Systems for ULSI Applications* (2000) 14.

[2] A. G. S. Filho, et al., *Proceedings of 16th Symposium Integrated Circuits and Systems Design* (2003) 99.

[3] T. W. Chen, et al., *IEEE International Symposium on Circuits and Systems, ISCAS* (2008) 2578.

[4] H. Shikano, et al., *Proceedings of the International Symposium on System-on-Chip, SoC* (2007) 7.

[5] Yitao Ma and Tadashi Shibata, *Proc. the 34th European Solid-State Circuits Conference, ESSIRC* (2008) Fringe P3.