Extended Abstracts of the 2006 International Conference on Solid State Devices and Materials, Yokohama, 2006, pp. 608-609

P-5-13L

Scalable Motion Estimation Core for Multimedia System-on-Chip Applications

Yeong-Kang Lai and Tian-En Hsieh

Department of Electrical Engineering, National Chung Hsing University
No.250, Kuo Kuang Road, Taichung, Taiwan
Phone: 886-4-22855268 E-mail: yklai@dragon.nchu.edu.tw

1. Introduction

Among various video compression techniques, the motion compensated hybrid coding is the most popular one and is adopted by several standards. Block matching algorithm for motion estimation is nowadays used in a wide variety of applications. The different performance requirements are needed for different real-time video applications. The performance requirements can be evaluated by some essential parameters such as the block size, the search area size, the frame size size, and the frame rate. To meet real-time video application, the number of PEs depends on the performance requirements.

Several dedicated hardware implementations have been realized for full-search-based block matching algorithms (FSBMA)[1]-[5]. Principally, these realizations are systolic arrays, laid out for a specific set of parameter values. Hence they usually offer only a limited flexibility or even no flexibility at all. Moreover, the parallel architecture with multiple PEs can efficiently increase throughput. However, the number of input pins, the difficulties on data addressing, and interconnection complexity between memory modules and PEs make it hard to implement. We propose a novel scalable architecture effectively to solve all these problems.

2. The Proposed VLSI Architecture

The procedure of a block-matching algorithm is to find the best matched displaced block from the previous frame F_{t-1}, within a search range, for each $N \times N$ block in the present frame F_t. A straightforward method, the full search, exhaustively matched all possible candidates to find the displacement (called motion vector) with a minimal distortion. As a criterion of distortion, the mean absolute difference (MAD) is calculated for each candidate location (u, v)

$$MAD(u,v) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \left| F_t(x+i, y+j) - F_{t-1}(x+i+u, y+j+v) \right|$$

where (x, y) is the coordinate of the upper-left pixel of the current block in F_t, and the values of u and v are limited to between $-p$ to $p-1$.

Fig.1 shows the scalable architecture to perform the FSBMA. To exploit the parallelism in FSBMA, the proposed architecture is consists of $H \times V$ identical on-chip memory modules (MMs) and $H \times V$ identical processing elements (PEs). The PEs are connected in a ringed fashion. Each PE is composed of an absolute difference unit, an accumulator, and a final-result latch in a pipelined fashion. According to the performance requirements of different real-time video applications, the number of the MMs and the PEs, i.e., $H \times V$, is determined by the following evaluation equation. Assume that the cycle time is T_p for the $N \times N$ current block with the search range of $-p$ to $p-1$.

$$T = \frac{\frac{W \times L}{N^2} \times (2p)^2 \times N^2 \times T_p}{H \times V} < \frac{1}{f_r}$$

Where T denotes the total block-matching time of all blocks in a frame (frame size: W×L). It should be smaller than frame period, 1/fr. The search area pixels to be computed are first stored in the MMs. The current block pixels are sequentially input and broadcasted to all PEs in a raster-can order. In most of multiple PEs designs, a fully connected interconnection is demanded between the multiple PEs and multiple and consumes larger routing space due to larger numbers of buses, multiplexers and tri-state buffers. To overcome the drawback, the adjacent PEs are connected in the horizontal rings and the vertical rings, and we arrange the data flow from the memory modules to the PEs and redistribute the PEs operations. This method allows the operations belonging to a candidate block to be performed by the several PEs. Then, for every clock cycle, PE propagates the accumulated partial-result to the adjacent PE to form the next partial-result of the block candidate. By use of the propagation of the accumulated partial-results, each of the PEs is only connected to one memory module. This eliminates the complicated interconnection and the switching circuitry between the PES and the memory modules. The upper-left H×V of the all candidate blocks are first computed. After 256 clock cycles, the MAD of each candidate block is produced in each PE and transferred to PE's final-result latch. These H×V latched MADs can then be sent to the minimum extractor to get a minimum MAD. During the comparisons, the PEs continue to perform the block–matchings of the next H×V candidate blocks. It does not consume extra cycles to fill up the pipeline operations.

3. Memory Interleaving Organization

The on-chip memory is used to reduce the load for chip I/O and memory system. The next problem is: how to provide all required data for the $H \times V$ PEs simultaneously. Our approach is to divide the memory into $H \times V$ memory modules, and to interleave the input pixels to these $H \times V$ mod-

ules. This architecture is based on some data management techniques: (1) On-chip memory configuration for data-reuse, (2) Memory interleaving organization for parallel data accesses, and (3) Propagation of the accumulated partial results for eliminating the interconnection overheads between the PEs and the MMs.

4. Performance Analysis

The chip layout is shown in Fig. 2. There are 16 PEs and 16 memory modules in the chip. With TSMC 0.18μm single poly six metal CMOS technology. It needs 18 input pads, 11 output pads, and 7 power pads. Simulations show that it has the capability to run up to 50 MHz, or 20 ns per cycle. Fig. 2 summarizes the characteristics of the chip.

5. Conclusion

A scalable PE-ringed architecture for FSBMA has been described. The number of processing elements (PES) is scalable according to the variable algorithm parameters and the performance required for different applications. A configuration of random-access on-chip memory modules solves the problems of chip I/O and memory bandwidth requirement. Input data are arranged in the memory modules by memory interleaving organization. Combined with a technique of the propagation of accumulated partial results, the interleaved memory module provides every PE with its required data simultaneously without introducing complicated interconnections and switching circuitry. In summary, the proposed architectures have the following desirable features: (1) low hardware cost, (2) high throughput rate, (3) low latency delay, (4) low 1/0 and memory bandwidth requirements, and (5) 100% computation efficiency.

References

[1] K. M. Yang, M. T. Sun, and L. Wu, "A family of VLSI designs for the motion compensation block matching algorithm," *IEEE Transcations on Circuits and Systems.*, vol. 36, pp. 1317-1325, Oct. 1989.

[2] C. H. Hsieh and T. P. Lin, "VLSI architecture for block-matching motion estimation algorithm," *IEEE Transcations on Circuits and Systems for Video Technology.*, vol. 2, pp. 169-175, June 1992.

[3] S. Kittitornkun and Y. H. Hu, "Frame-level pipelined motion estimation array processor," *IEEE Transcations on Circuits and Systems.*, vol. 11, no. 2, pp. 248-251, Feb. 2001.

[4] N. Roma and L. Sousa, "Efficient and Configurable Full-Search Block-Matching Processors," *IEEE Transcations on Circuits and Systems.*, vol. 12, no. 12, pp. 1160-1167, Dec. 2002.

[5] Y. K. Lai, L. G. Chen, T. H. Tsai, and P. C. Wu, "A novel scalable architecture with memory interleaving organization for full search block-matching algorithm," *IEEE International Symposium on Circuits and Systems.*, vol. 2, pp. 1229-1232, June 1997.

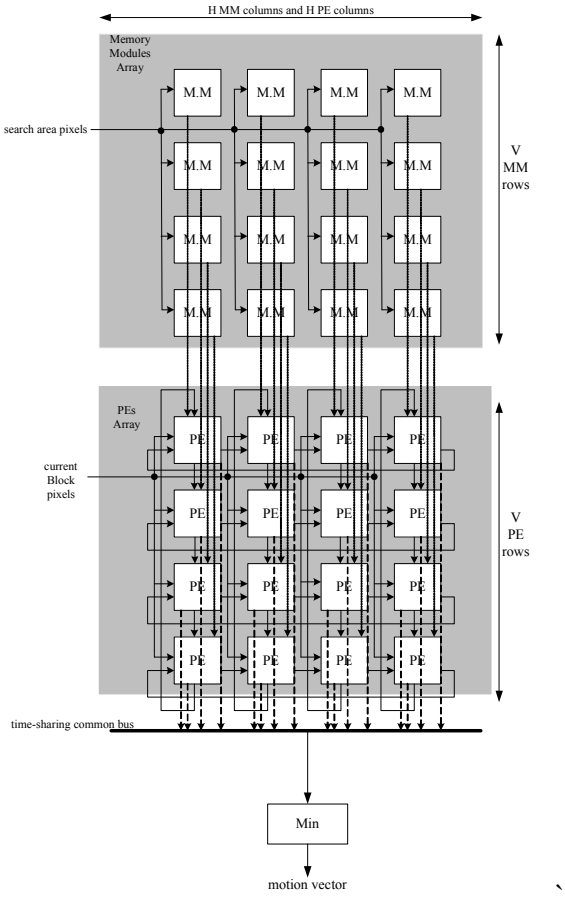

Fig.1 The scalable PE-ringed architecture

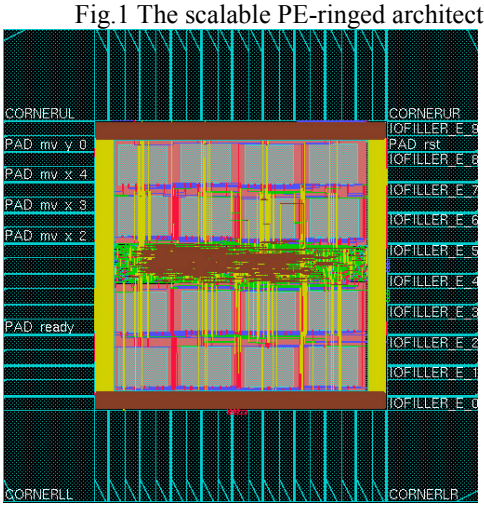

Fig. 2. Chip layout

Table 1. Chip features

Technology	**TSMC 0.18um 1P6M**
Chip Size	**1488807.956 um^2**
Memory Size	**32x8**
Gate Count(Nand2)	**33503**
Working Frequency	**50MHz**
Voltage	**1.8v**
Power	**13.1487mW**
I/O PAD	**36**