A 211 GOPS/W Dual-Mode Real-time Object Recognition Processor with Network-on-Chip

Kwanho Kim, Joo-Young Kim, Seungjin Lee, Minsu Kim, and Hoi-Jun Yoo
Department of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
kkh82@eeinfo.kaist.ac.kr

***Abstract*—This paper presents a 211 GOPS/W real-time object recognition processor with network-on-chip (NoC). The chip integrates 8 linearly connected SIMD clusters with 8 4-way VLIW processing elements (PEs) per cluster. The SIMD/MIMD dual-mode object recognition processor exploits both data-level and object-level parallelism based on the NoC configuration. The 8-way SIMD PE cluster is optimized for data-intensive object recognition tasks. Packet-based power management scheme is employed for low power consumption. The proposed processor takes 36mm^2 in 0.13µm CMOS process and achieves a peak performance of 96GOPS at 200MHz with 392mW power consumption.**

I. INTRODUCTION

Recently, intelligent vision processing such as object recognition and video analysis has been widely used in various applications such as mobile robot navigation, automotive vehicle control, video surveillance, and natural human-machine interfaces. Such vision applications require huge computational power and real-time response under the low power constraint, especially for mobile devices. Programmability is also needed to cope with a wide variety of applications and recognition targets.

Object recognition involves complex image processing tasks which can be classified into several stages of processing with different computational characteristics. In low-level processing (e.g. image filtering, feature extraction), simple arithmetic operations are performed on a 2-D image array of pixels. On the contrary, processing at higher levels is irregular and performed on objects that are groups of features extracted at the lower level. Because object recognition requires huge computation power on each stage, general-purpose processor cannot achieve a real-time performance due to its sequential processing. Many vision processors were previously reported based on massively parallel SIMD paradigm with a number of processing elements (PEs) for data-level parallelism [1-2]. However, these processors focus on only the low-level image processing operations like image filtering and thus they are not suitable for object-level parallelism, which occupies a relatively large portion on higher level vision applications such as object recognition. A multiple-instruction multiple-data (MIMD) multi-processor was presented with Network-on-Chip (NoC) to exploit task-level parallelism [3]. However, it cannot reach a real-time performance due to its limited computing power and required complex data synchronization mechanism.

To overcome the computational complexity of the object recognition, visual attention based object recognition algorithm has been developed as shown in Fig. 1 [4]. Visual attention is the ability of the human visual system to rapidly select the most salient part of the image [5]. By the visual attention mechanism, the image region of interests is selected in a pre-attentive phase. Then, next visual processing such as key-point extraction, feature vector generation and matching focus on only the pre-selected image in a post-attentive phase. Therefore, computation cost reduction can be obtained by drastically reducing the amount of the image data to be processed on higher-level image processing tasks.

In this paper, a power efficient dual-mode real-time object recognition processor is presented for the attention-based object recognition applications. The proposed processor which integrates 8 linearly connected SIMD PE clusters can be configured into a SIMD or MIMD mode by adaptively selecting circuit or packet switching of the NoC in order to exploit both data-level and object-level parallelism. The 8-way SIMD PE cluster with 8 4-way very long instruction word (VLIW) PEs is specialized for object recognition tasks. A packet-based power management is employed for low power consumption. As a result, the object recognition processor achieves a peak performance of 96GOPS at 200MHz with 392mW power consumption while object recognition is running at 22 frames/sec.

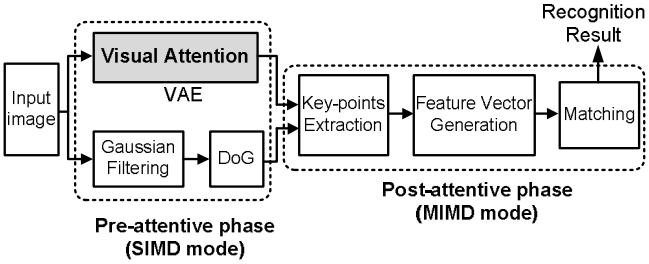

Figure 1. Attention-based object recognition

978-1-4244-2362-0/08/$25.00 ©2008 IEEE.

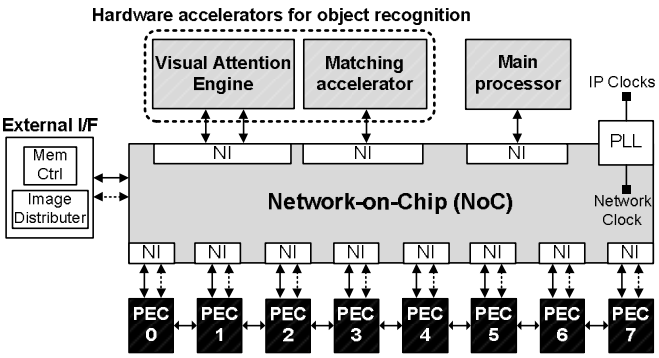

Figure 2. System architecture of the object recognition processor

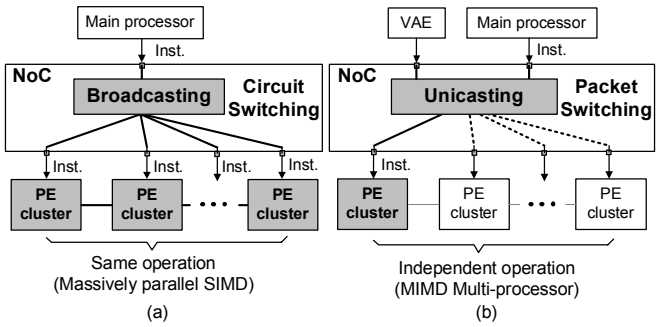

Figure 3. Dual-mode operations based on the NoC configuration

II. System Architecture

A. System Operation

Fig. 2 shows the overall architecture of the proposed NoC-based object recognition processor, which consists of a main processor, a visual attention engine (VAE), a matching accelerator (MA), 8 PE Clusters (PECs) and an external interface. The ARM10-compatible 32-bit main processor controls the overall system operations. The VAE, an 80x60 digital cellular neural network, rapidly detects the salient image regions on the sub-sampled image (80x60 pixels) by neural network algorithms like contour and saliency map extraction [4]. The 8 linearly connected PECs perform data-intensive image processing applications such as image gradients and histogram calculations for further analysis of the salient image parts (i.e., the objects) provided by the VAE. The MA boosts nearest neighbor search to obtain a final recognition result in real-time. The DMA-like external interface distributes automatically the corresponding image data to each PEC to reduce system overhead. Initially, 2-D image plane is equally divided into 8 PECs according to the image size specified by the main processor. Each core is connected to the NoC via a network interface (NI).

B. Dual-mode Configuration

The attention-based object recognition applications require a wide range of parallelism: data-level parallelism for the entire image as a pre-attentive phase and object-level parallelism for only salient image regions selected by the VAE as a post-attentive phase (See Fig. 1). To address the above requirements, the proposed object recognition processor has dual-mode configuration. According to the NoC configuration, the system has two different operation modes as shown in Fig. 3: SIMD and MIMD mode. In a circuit switching NoC, the main processor broadcasts instruction and data to all PE array. In this mode, the system exploits massively parallel SIMD operation for image pre-processing, achieving the peak performance of 96 GOPS at 200 MHz. On the contrary, in a packet switching NoC, each PEC is responsible for the objects, each of which contains image data around the extracted key-points. In the MIMD mode, the 8 PECs operate independently in parallel for object-parallel processing.

It takes about a few tens of cycles to change the NoC configuration depending on the network traffic status due to circuit establishment and release time overhead for the circuit switching NoC. For object recognition application, however, the operation mode conversion occurs only twice during the recognition period of 1-frame image: SIMD to MIMD conversion after the pre-processing stage and MIMD to SIMD conversion after completing the recognition. Therefore, such a dual-mode architecture is suitable for object recognition with negligible impact on the overall system performance.

III. PE Cluster Design

A. Overall Architecture

The PEC is a SIMD processor array designed to accelerate image processing tasks. Fig. 4 shows the block diagram of the PEC. It contains 8 linearly-connected PEs controlled by a cluster controller, a cluster processing unit (CLPU), 20 kB local shared memory (LSM), a LSM controller, and a PE load/store unit. The 8 PEs operate in a SIMD fashion and perform image processing operations in a column-parallel (or row-parallel) manner. The CLPU, which consists of an accumulator and an 8-input comparator, generates a single scalar result from the parallel output processed by the PE array. The LSM is used as on-chip frame memory or local memory for each PEC to store the input or processed image data and objects. A single-port 128-bit wide SRAM is used for the LSM to reduce area overhead. The LSM provides a single-cycle access and is shared between the PE load/store unit, the LSM controller and the CLPU. Arbitration for the LSM is performed on a cycle-by-cycle basis to improve the LSM utilization. The LSM controller is

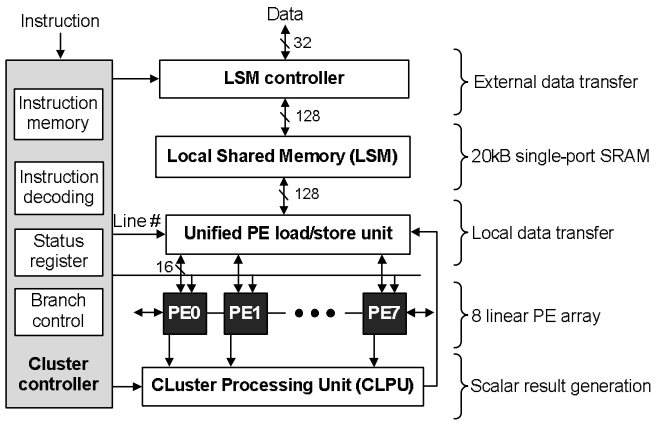

Figure 4. Block diagram of the PEC

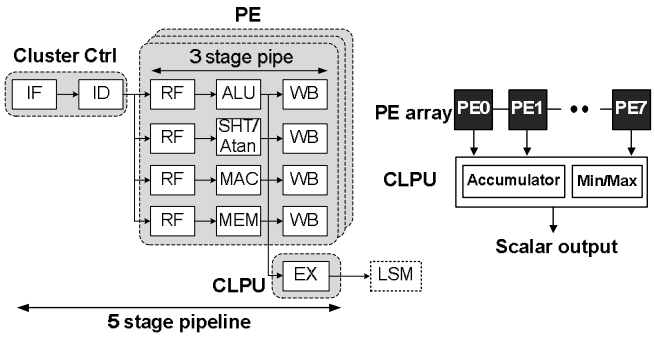

Figure 5. Tightly coupled 5-stage pipeline of the PEC

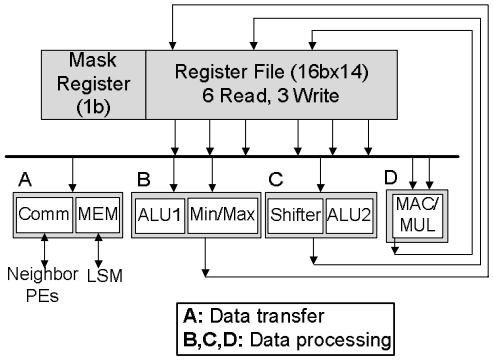

Figure 6. Block diagram of 4-way VLIW PE

responsible for data transfer between external memory or other PECs and the LSM while the PE load/store unit can access the LSM only for local data transfer. The LSM controller, which is an independent processing unit optimized for data transfer like the DMA engine, enables the data transfers in parallel with PE execution to hide excessive external memory latency. In addition, due to the simple control circuit in the SIMD architecture, the cluster controller including 2 KB instruction memory occupies only 6% of the total PEC area, which results in high computation efficiency.

B. Tightly coupled Pipeline Structure

Fig. 5 shows the 5-stage pipeline architecture of the PEC. The cluster controller, the 3-stage pipelined PE array, and the CLPU are tightly coupled to maintain 1-cycle throughput for all operations. Especially, the tightly coupled PE array and CLPU architecture achieves single-cycle execution for statistical image processing tasks (e.g. histogram calculations) where an input image is transformed into a scalar or vector data, while the massively parallel SIMD processors [1,2] require sequential operations on a line-by-line basis to obtain the same result due to the absence of the CLPU-like vector-to-scalar processing unit. Such an architecture is suitable for object recognition because histogram calculations is the essential operation for key-point descriptor generation in the object recognition task [6].

C. 4-WAY VLIW PE

Each PE utilizes 4-way VLIW architecture to execute up to 4 instructions in a single cycle as shown in Fig. 6: three instructions for data processing (B,C,D) and one instruction for data transfer (A). It consists of two 16-bit ALUs, a shifter,

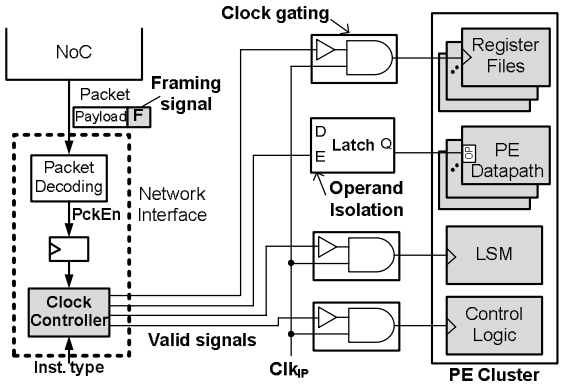

Figure 7. Packet-based power management

a multiplier, a 16-bit 9-port register file and a 1-bit mask register for supporting a conditional execution in a SIMD PE array. All PE instructions have single-cycle execution except 16-bit multiply-accumulate (MAC) operation, which has a two-cycle latency. The 16-bit datapath units of the PE can be configured to execute two 8-bit operations in parallel for gray-scale image processing. The resulting peak performance is 89.6 GOPS (8-bit fixed point) for 64 PE array and 6.4 GOPS for 8 CLPUs at 200MHz. The left and right neighbor PE registers can be directly accessed in a single-cycle using the linearly connected PE array for efficient inter-PE communication, which is one of the most frequently used operations for neighborhood image processing tasks such as image filtering. Meanwhile, memory access patterns are well predictable for such low-level image processing tasks due to the characteristics of regular and pre-defined data accesses. The 4-way VLIW PEs allow PEC software to pre-fetch the needed data in advance without performance loss by executing data transfer and processing instructions concurrently.

IV. PACKET-BASED POWER MANAGEMENT

The modular and point-to-point NoC approach makes it easy to manage the overall system by decoupling computation of IPs from inter-IP communication, which enables efficient power management techniques compared to the bus-based system. For low power consumption, our chip performs packet-based power management at the IP level as shown in Fig. 7. Each PE cluster is individually enabled or disabled according to the framing signal of the packet to cut the power of inactive IPs. The valid signals generated by the network interface wake up the appropriate blocks within the IP only when incoming packet arrives. 4 clock domains of the PE cluster are individually controlled based on the issued instruction type. During the image data transfer phase for which only the LSM controller needs to be activated, the clock signals of the PE register files are gated-off and operand isolation to the PE datapath prevents unnecessary signal transitions to reduce power consumption. Since the PE datapath and register files occupy about 62% of the total power consumption, the power reduction up to 27% is achieved when the object recognition application is running. The packet-based power management scheme can be generally extended to a NoC-based multi-core system for the IP-level power control.

V. Implementation Result

The proposed object recognition processor is fabricated in a 0.13 μm 1-poly 8-metal standard CMOS logic process, and its die area takes 6 x 6 mm^2 including 1.9M gate count and 228kB on-chip SRAM. The chip micrograph and evaluation board are shown in Fig. 8 and Table I summarizes the chip features. Operating frequency of the chip is 200MHz for the IPs and 400MHz for the NoC. The power consumption is about 392mW (excluding the VAE and MA) at 1.2V power supply while object recognition application is running at 22 frames/sec.

Fig. 9 shows the comparison with the previously reported vision processors in terms of power efficiency [1-3,7]. To normalize the value, GOPS/W and nJ/W are adopted as a performance index. As a result, the chip achieves up to 4.3 times higher GOPS/W in case of 8-bit fixed-point operation and energy per pixel reduction up to 70% is obtained for object recognition task. Fig. 10 shows the performance evaluation when the object recognition application is running on the chip. In this example, the VAE performs a saliency map extraction as attention cues and 50 objects are used as a database for matching. In the SIMD mode, the VAE and 8 PECs take 14.4 ms to complete saliency map extraction and difference-of-Gaussian filtering while exploiting data-level parallelism of the 64 PE array. In the MIMD mode, higher-level vision tasks such as feature vector generation and matching are performed on objects extracted at the lower level stage while exploiting object-level parallelism. As a result, the chip achieves 22 frames/sec recognition speed without degradation of recognition rate, which is sufficient for real-time operation.

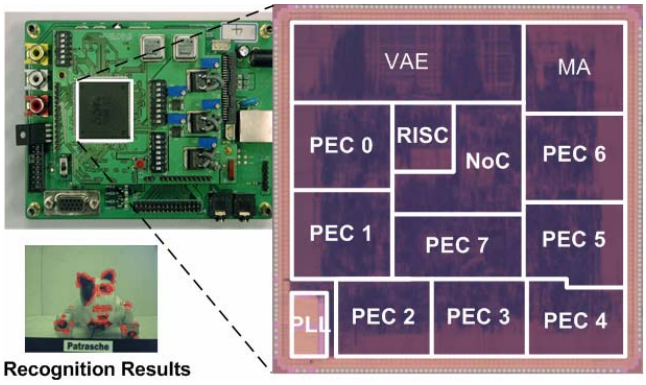

Figure 8. Chip micrograph and evaluation board

TABLE I. Chip features

Process	0.13um 1P 8M CMOS technology
Die Size	6mm x 6mm
Power Supply	1.2V for core, 2.5V for I/O
Operating Frequency	400MHz for NoC 200MHz for IPs
# of TRs (gates, memory)	1.9M gates, 228kB SRAM
Power Consumption	< 392mW (for full applications)
Peak Performance	96GOPS (for 8 PE clusters)
Object Recognition Rate	22 frame/sec @ 320x240 image

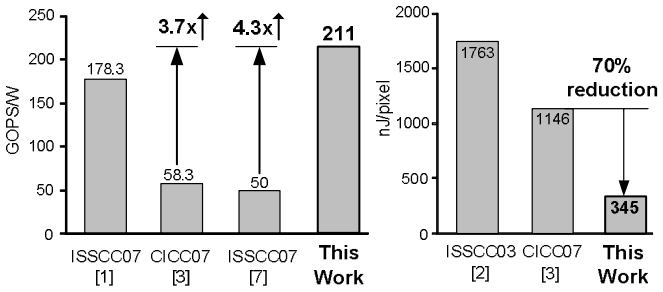

Figure 9. Power efficiency comparison

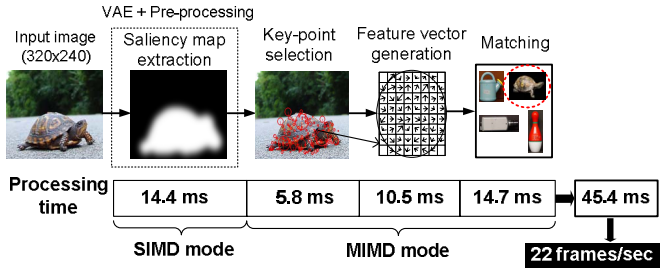

Figure 10. Performance evaluation

VI. Conclusion

A high performance and power efficient real-time object recognition processor is proposed for intelligent vision applications. The processor supports both the SIMD and MIMD computing modes based on the NoC configuration. The 8 linearly connected PE clusters optimized for object recognition tasks enable high performance image processing while achieving the low power consumption with the help of the packet-based power management. The chip achieves a peak performance of 96GOPS at 200MHz while dissipating 392mW from 1.2V power supply. The evaluation board with the fabricated chip demonstrates the real-time object recognition for intelligent mobile robot vision system.

REFERENCE

[1] A. Abbo, et al., "XETAL-II: A 107 GOPS, 600mW Massively-Parallel Processor for Video Scene Analysis," *ISSCC Dig. of Tech. Papers*, pp. 270-271, 2007.

[2] S. Kyo, et al., "A 51.2-GOPS Scalable Video Recognition Processor for Intelligent Cruise Control Based on a Linear Array of 128 Four-Way VLIW Processing Elements", *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1992-2000, Nov. 2003.

[3] D. Kim, et al., "An 81.6 GOPS Object Recognition Processor Based on NoC and Visual Image Processing Memory," *Proc. of CICC,* pp. 443-446, 2007.

[4] K. Kim, et al., "A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-inspired Visual Attention Engine," *ISSCC Dig. of Tech. Papers*, pp. 523-524, 2008.

[5] L. Itti, et al., "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, Nov. 1998.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints", *Int. J. Comput. Vis.*, vol. 60, no. 2, pp.91-110, Nov. 2004.

[7] B. Khailany, et al., "A Programmable 512 GOPS Stream Processor for Signal, Image, and Video Processing," *ISSCC Dig. of Tech. Papers*, pp. 272-273, 2007.