A 76.8 GB/s 46 mW Low-latency Network-on-Chip for Real-time Object Recognition Processor

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*—A 76.8 GB/s 46 mW low-latency network-on-chip (NoC) provides a communication platform for a real-time object recognition processor. The tree-based topology NoC with three crossbar switches is designed for low-latency by adopting dual-channel and adaptive switching. The NoC can be dynamically configured to exploit both data-level and object-level parallelism on the object recognition processor. FLIT-level clock gating and packet-based power management scheme are employed for low power consumption. The NoC is implemented in 0.13μm CMOS process and provides 76.8 GB/s aggregated bandwidth at 400MHz with 2-clock cycle latency while dissipating 46 mW at 1.2 V.**

I. INTRODUCTION

Object recognition has been widely used in various applications such as mobile robot navigation, autonomous vehicle control, video surveillance, and natural human-machine interfaces. These applications require huge computational power and real-time response under the low power constraint, especially for mobile devices. In addition, programmability is needed to cope with a wide variety of applications and recognition targets.

Recently, visual attention based object recognition algorithm has been developed to overcome the computational complexity of the object recognition [1]. Visual attention is the ability of the human visual system to rapidly select the most salient part of the image [2]. By the visual attention mechanism, the image region of interests is pre-selected, thus further visual processing focus on only the pre-selected objects. 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. For hardware implementation of such vision system, a number of processing elements (PEs) are needed to handle compute-intensive image processing. A large amount of data transactions among them are also required to facilitate object-level parallel processing. Therefore, the network-on-chip (NoC) based solution is suitable to achieve extensive bandwidth among PEs.

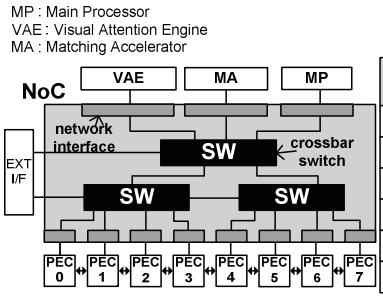
Several vision processors have been reported for object recognition. Massively parallel SIMD processors with a number of PEs were presented to exploit data-level parallelism in a 2-D image array of pixels [3,4]. However, these processors focus on only the low-level image processing operations like image filtering and thus they are not suitable for object-level parallelism, which is essential for the object recognition. A multiple-instruction multiple-data (MIMD) multi-processor was presented with the NoC to exploit task-level parallelism [5]. However, it cannot reach a real-time performance due to its limited computing power and required complex data synchronization mechanism.

In this paper, a low-latency and energy-efficient NoC is presented for the bio-inspired object recognition processor [1]. The NoC is dynamically configured to exploit both data-level and object-level parallelism for the attention-based object recognition. The tree-based topology NoC with three crossbar switches supports dual-channel and adaptive switching techniques for low-latency and high energy efficiency. FLIT-level clock gating and packet-based power management scheme are used for low power consumption. As a result, the NoC provides 76.8 GB/s aggregated bandwidth at 400 MHz with 2-clock cycle latency while dissipating 46 mW at 1.2 V.

II. NOC REQUIREMENTS

A. Traffic Characteristics

Fig. 1 shows the overall architecture of the proposed NoC-based object recognition processor with major traffic flows. The tree-based topology NoC with three crossbar switches interconnects 12 IPs: a main processor, VAE [6], a matching accelerator, 8 PE Clusters (PECs) and an external interface. Although regular topology NoC (e.g. mesh, torus) has been



Major traffic flows

Source	Destination	Traffic characteristics
MP	PECs	Broadcast
MP	VAE	No bursty
MP	MA	No bursty
PEC	PEC	Bursty
PEC	EXT I/F	Bursty
VAE	PEC	No bursty

Figure 1. Object recognition processor with major traffic flows

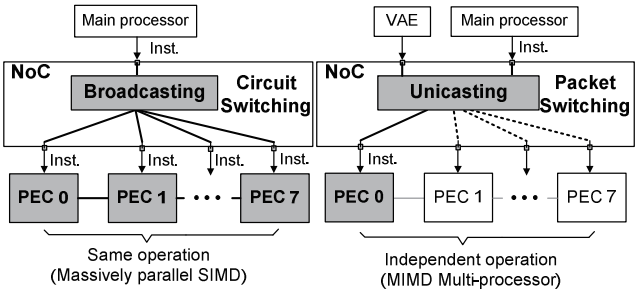


Figure 2. Dual-mode configuration

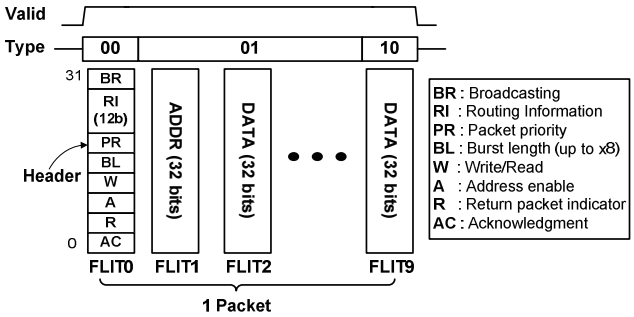


Figure 3. NoC packet format

widely used for its better scalability, the tree-based customized NoC achieves lower latency and power than a 2-D mesh. The object recognition processor has two different kinds of traffic flows: bursty traffics for image data transfer and non-bursty traffics for control and synchronization between IPs. The traffic among PECs mainly consists of high-bandwidth image data so that the PEC performs data-intensive image processing applications. The external interface also communicates with the PEC to transfer bursty image data. The main processor generates the non-bursty traffics to control the overall system operations. The VAE transfers a set of salient image regions to the corresponding PEC.

B. NoC Requirements

The object recognition processor places a number of requirements on the NoC. Each PEC, operating at 200 MHz, is capable of 12.8 Gb/s of aggregated throughput, thus the NoC should provide sufficient bandwidth to deal with large image data transfer between the PECs. The low-latency is also one of the most important NoC features for real-time operation. The attention-based object recognition requires 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. To address the above requirements, the NoC should support packet broadcasting to all PECs for SIMD operation as well as a conventional packet switching for MIMD operation. In this work, the dynamically configured NoC determines the system operation modes as shown in Fig. 2: 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 125 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. Because the traffic characteristics among the PECs are dynamic and variable depending on the number of key-points and their location on the image, the NoC-based solution is more appropriate than the multi-layer bus.

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 key-points extraction 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. LOW-LATENCY NOC

A. NoC Protocol

Fig. 3 shows the NoC packet format. A wormhole switching, where each packet is divided into a few 32-bit FLITs (FLow control unIT) with additional 3-bit control signals, is employed to reduce buffer requirements. Header FLIT contains 4-bit burst length information for burst packet transaction up to 256-bit (8x32-bit) packets and 2-bit priority information for quality-of-service (QoS) control. A handshaking protocol is supported for reliable transmission by using an acknowledgement request. The packet length is determined by the burst length and maximum 10 FLITS are possible. Deterministic source routing scheme is used for simple hardware implementation. Circuit and packet switching are adaptively selected for a specific route path from the main processor to the PECs in order to support dual-mode configuration. A 1-bit sideband back-pressure signal is used for the flow control in the NoC. The back-pressure signal is asserted to stop the packet transmission when buffer overflow occurs, or when destination PE cannot provide the required service.

B. Low-latency Crossbar switch

Fig. 4 shows the block diagram of the proposed low-latency crossbar switch. A dual-channel crossbar switch is designed for low-latency. At port 0, the switch supports both circuit and packet switching adaptively according to the system operation mode. In circuit switching mode, burst packet can be broadcasted to all PECs by bypassing the queuing buffers and arbiter, resulting in reduced delay and energy dissipation. An input driver, which consists of 3 tri-state buffers coupled in parallel, dynamically controls its drive strength according to the switching mode for reliable packet transmission. At port 1 through 6, image-express channel is added to normal packet channel in order to reduce latency of the return packets transferring from slave IPs. The return packet latency seriously affects the overall system performance because the PEC with in-order execution stalls until the return packet arrives. Incoming return packets, detected by a pre-route unit, are ejected immediately after 1 flit buffer through an additional

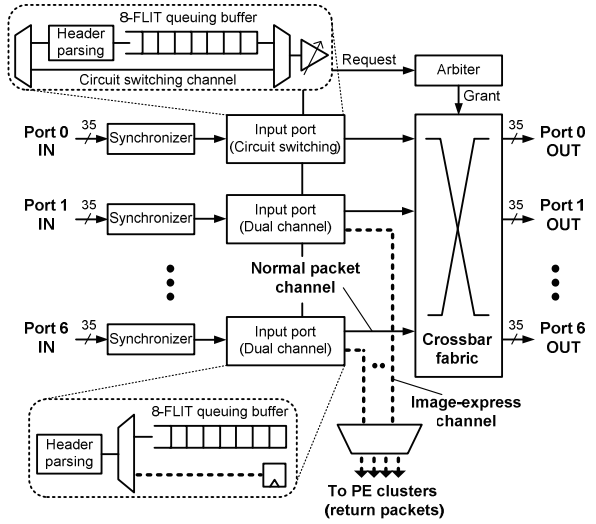


Figure 4. Block diagram of the proposed crossbar switch

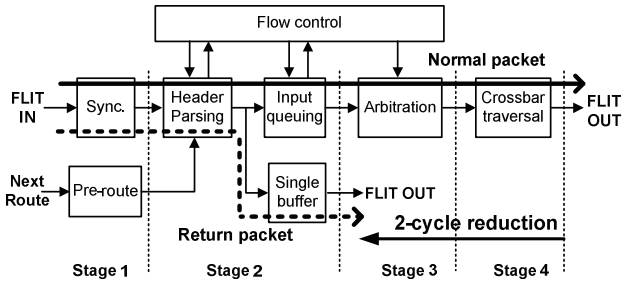


Figure 5. Crossbar switch pipeline

image-express channel. This mechanism saves 2 pipeline stages of the switch by eliminating unnecessary packet queuing, arbitration, and crossbar fabric traversal.

Fig. 5 shows the 4-stage low-latency crossbar switch pipeline. Incoming return packets are ejected 2-cycle earlier than normal packets without any flow control. Because return packets are mostly burst packets, this scheme is more effective. The crossbar switch with the dual-channel does not store the return packets at queuing buffers. They are directly injected into the network without any suppression by the back-pressure flow control, which leads to a significant performance improvement over a conventional crossbar switch [7]. As a result, 26% latency reduction is obtained with only 6% area overhead compared to the conventional crossbar switch [7] while various image processing applications are running on the NoC-based system.

C. FIFO Synchronizer

A first-in-first-out (FIFO) based synchronizer (See Fig. 6) is designed to interface between the IPs and the NoC with independent clock frequencies and phases. Without global synchronization, packet transmission is performed by source synchronous scheme in which a strobe signal is transmitted along with the packet data [8]. A 4 FLITs depth FIFO captures the incoming FLIT using the delayed strobe signal. Detection of the full or empty status is accomplished using the FIFO write and read pointers to avoid FIFO overflow or underflow,

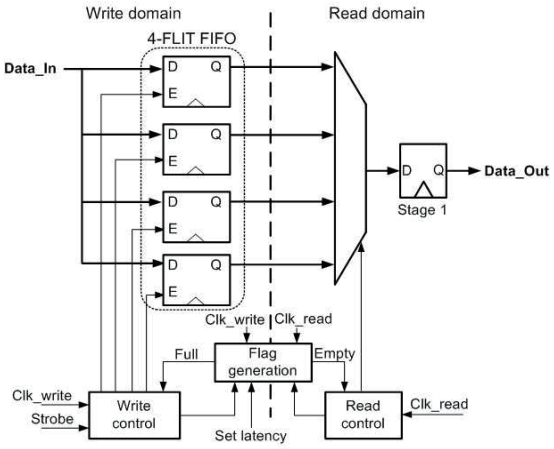


Figure 6. FIFO-based synchronizer

respectively. The synchronizer is placed at the first stage of the crossbar switch pipeline.

IV. Low Power Techniques

A. FLIT-level Clock Gating

For low power consumption, we apply FLIT-level fine-grained clock gating to the crossbar switch. Only the required packet routing path is activated on a per-port basis. A power control unit that monitors an incoming packet header is always turned on and the output port number is encoded to control the NoC clock signal. The clock gating signals are generated on all pipelining stages of the crossbar switch in a pipelined manner. Only queuing buffer at port N, arbiter and crossbar fabric at port M are enabled when a FLIT is transferred from an input port N to an output port M. Since queuing buffers, built using a number of flip-flops, are the most power-consuming unit in the NoC, the FLIT-level power management can reduce NoC power consumption by 32% without degradation of throughput and latency.

B. Packet-based IP-level 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. Each PEC is individually enabled or disabled according to the framing signal of the packet to reduce 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 PEC are individually controlled based on the issued instruction type. During the image data transfer phase for which only the local shared memory 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 Results

The proposed NoC-based object recognition processor is fabricated in a 0.13µm 8-metal CMOS process. Its die area takes 6x6 mm^2 including 1.9 M gate count and 228 kB on-chip SRAM, where the NoC contains 22k gate count. The chip micrograph and the evaluation board are shown in Fig. 7. Operating frequency of the chip is 200 MHz for the IPs and 400 MHz for the NoC. The NoC provides 4.8 GB/s communication bandwidth for each PEC and 76.8 GB/s aggregated bandwidth (60.8 GB/s for normal channel and 16 GB/s for image-express channel) at 400 MHz. The total chip power consumption is about 583 mW at 1.2 V power supply while object recognition application is running at 22 frames/sec, where the NoC accounts for 46 mW. Fig. 8 shows the power and area breakdown of the object recognition processor. The NoC consumes 9% of the die area and 8% of the power consumption, which means that the NoC cost is amortized over the processing units.

Measured waveforms (See Fig. 9) show the low-latency return packet transmission by the crossbar switch with dual-channel when the NoC operates at 200 MHz. The incoming return packets are transferred to the output port 2-clock cycles earlier than normal packets. With the help of the low-latency NoC platform and packet-based power management, the chip achieves up to 8.3 times higher power efficiency than previous works [3-5] in terms of GOPS/W.

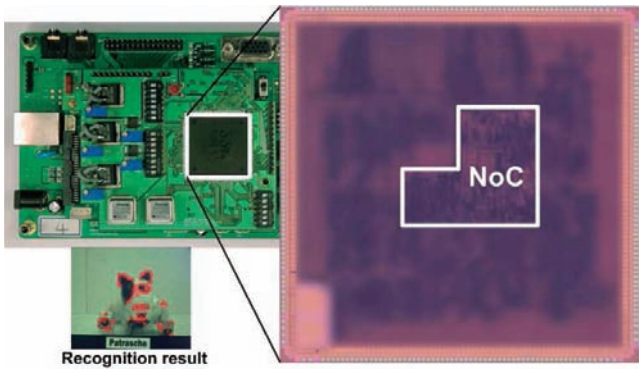


Figure 7. Implementation results

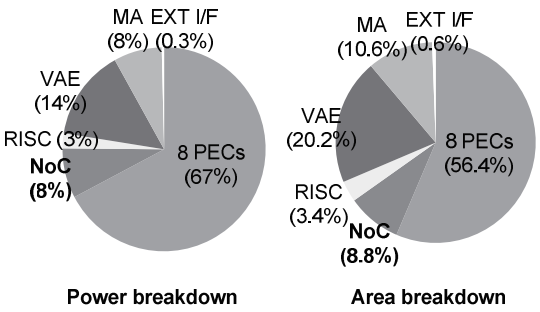


Figure 8. Power and area breakdown

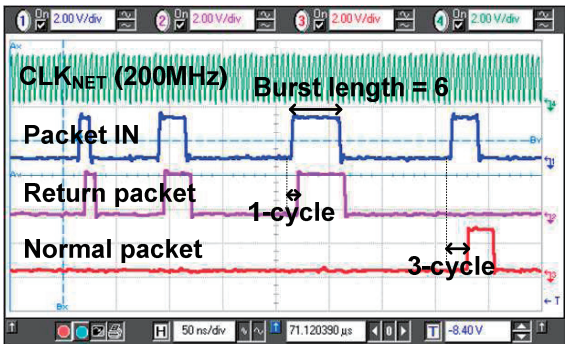


Figure 9. Measured waveforms of the NoC packet signals

VI. Conclusion

A low-latency and energy-efficient NoC is proposed as a communication platform for the real-time object recognition processor. The tree-based topology NoC with three crossbar switches is optimized for low-latency and energy efficiency by adopting dual-channel and adaptive circuit/packet switching techniques. The NoC can be dynamically configured to exploit both data-level and object-level parallelism on the object recognition processor. FLIT-level clock gating and packet-based power management scheme are employed for low power consumption. The NoC is fabricated in a 0.13µm CMOS process and provides 76.8 GB/s aggregated bandwidth at 400 MHz with 2-clock cycle latency while dissipating 46 mW at 1.2 V.

Reference

[1] 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. 308-615, 2008.

[2] 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.

[3] A. Abbo, et al., "XETAL-II: A 107 GOPS, 600mW Massively-Parallel Processor for Video Scene Analysis," *ISSCC Dig. of Tech. Papers*, pp. 270-602, 2007.

[4] 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", *ISSCC Dig. of Tech. Papers*, pp. 48-477, 2003.

[5] 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.

[6] S. Lee, et al., "The Brain Mimicking Visual Attention Engine: An 80x60 Digital Cellular Neural Network for Rapid Global Feature Extraction", *IEEE Symposium on VLSI*, pp. 26-27, 2008.

[7] D. Kim, et al., "Implementation of Memory-Centric NoC for 81.6 GOPS Object Recognition Processor", *IEEE Asian Solid State Circuits Conf.*, pp.47-50, 2007.

[8] K. Lee, et al., "Low-Power Networks-on-Chip for High-Performance SoC Design," *IEEE Trans. VLSI Systems*, vol. 14, no.2, pp.148-160, Feb. 2006.