16.2 A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-inspired Visual-Attention Engine

Kwanho Kim, Seungjin Lee, Joo-Young Kim, Minsu Kim,
Donghyun Kim, Jeong-Ho Woo, Hoi-Jun Yoo

KAIST, Daejeon, Korea

The visual attention algorithm of the human visual system [1] is utilized to reduce the complexity of object recognition by decreasing the amount of image data to be processed. As Fig. 16.2.1 illustrates, salient parts of a scene are roughly selected by the visual attention mechanism in advance so that next visual processing can focus on only the pre-selected objects to reduce the computational cost by 42%. For such a bio-inspired vision system, multiple parallel processing elements (PEs) and a large amount of data transactions among them are required. Recently, a massively parallel processor was presented for data-level parallelism, but it was not suitable for object-parallel processing due to limited interconnections between PEs [2]. A Network-on-Chip (NoC) is applied to achieve extensive communication bandwidth required for parallel computing. A 125GOPS NoC-based parallel processor with a bio-inspired visual attention engine (VAE) exploits both data and object-level parallelism while dissipating 583mW by packet-based power management. The use of more PEs, VAE, and low latency NoC enables higher performance and power efficiency over the previous design [4].

Fig. 16.2.2 shows the block diagram of a NoC-based parallel processor consisting of 12 IPs: a main processor, 8 PE clusters (PECs), VAE, a matching accelerator (MA), and an external interface. The ARM10-compatible 32b main processor controls the overall system operations. The VAE detects the feature points on the entire image by neural network algorithms like contour extraction. The 8 PECs perform data-intensive image processing applications such as filtering and histogram calculations. The MA accelerates 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. Each core is connected to the NoC via a network interface (NI).

The PEC contains 8 linearly-connected PEs controlled by a cluster controller, 20kB local shared memory (LSM), and a LSM controller. Each PE is a 3-stage pipelined 16b datapath with a 9-port register file and can execute up to 4 instructions in parallel, all executed in one cycle except 16b multiply-accumulate (MAC) operations. The LSM controller is responsible for data transfer between external memory and LSM, which enables the data transfers in parallel with PE execution to hide memory latency. The NoC can configure the system dynamically into different operation modes. In a circuit switched NoC, the main processor broadcasts instruction and data to a 64 PE array in order to exploit highly-parallel SIMD functions, achieving the peak performance of 96GOPS at 200MHz. To facilitate object-parallelism, the 8 PECs operate independently in parallel as a packet switched NoC.

The tree-based NoC with 3 star-connected crossbar switches is used for low latency and power [3]. The NoC protocol supports burst packet transactions of up to 8 packets and handshaking for reliable transmission. A 7×7 crossbar switch of Fig. 16.2.3 is optimized for low latency and energy efficiency with two key features: dual channel (port 1-6) and adaptive switching (port 0). Packet latency for transferring read data from slave IPs seriously affects the overall system performance because the PEC stalls until the return packet arrives. Incoming return packets, detected by a pre-route unit, are ejected immediately after a 1 flit buffer through an additional image-express channel. This mechanism saves 2 pipe-stages of the switch by eliminating unnecessary packet queuing, arbitration, and switch fabric traversal. The crossbar switch also supports both circuit and packet switching adaptively according to the system operation modes. In circuit switching mode, burst packet can be broadcasted to all PECs by bypassing queuing buffers and arbiter, resulting in reduced delay and energy dissipation. An input driver at port 0 dynamically controls its drive strength based on the output load associated with the switching mode for reliable packet transmission. Combined application of both techniques achieves 43% latency and 48% energy reduction with only 6% area overhead compared to a conventional crossbar switch [4], while image processing applications are running on the NoC-based system.

Fig. 16.2.4 shows the block diagram of the VAE, which is composed of 4 arrays of 20×60 cells, 120 visual PEs (VPEs) shared between the cell arrays, and a controller. An 80×60 shift register array, distributed among the cells, eliminates data communication overhead in convolution operations of arbitrary kernel size and shape. Its cellular neural network based architecture [5] emulates human brains to accelerate visual attention algorithms like contour, texture, and motion extraction. Each cell, corresponding to a pixel in an image, contains 4 of 8b 6T SRAM-based registers for storing intermediate data, and a 4-directional shift register. In order to minimize cell area, the shift register is implemented using NMOS dynamic logic. Also, since all cells shift in the same direction, one channel is sufficient for 2-way communication between two neighboring cells, reducing the wiring complexity. As a result, the full-custom designed cell occupies a compact area of 496μm^2, achieving a reduction of 40% over the static MUX-based design. The VPEs, operating in SIMD mode, are capable of 1 cycle MAC operation and employ 3 stage pipelines that consist of read, execute, and write. To facilitate 1 cycle throughput, read and write of cell data is sequentially executed within one cycle using a self-timed circuit, as shown in Fig. 16.2.5. The resulting peak performance of the 120 VPEs is 24GOPS at 200MHz. The VAE takes only 18μs to complete contour extraction: a 10^3 times improvement over a 2.66GHz Intel Core 2 processor.

Packet-based fine-grained clock gating is used to reduce NoC power consumption as shown in Fig. 16.2.6a. 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 generates clock gating signals on all pipelining stages of the crossbar switch in a pipelined manner. The packet-level power management reduces NoC power consumption by 32% without degradation of throughput and latency. In addition, each IP is individually enabled according to the framing signal of the packet to cut the power of inactive IPs. The valid signals generated by the NI wake up the appropriate blocks within the IP only when incoming packet arrives (Fig. 16.2.6b).

The chip is fabricated using 0.13μm CMOS technology and its die size is 36mm^2 including 1.9M gates and 228kB SRAM. Power consumption is 583mW at 200MHz and 400MHz NoC with 1.2V power supply (25°C), while object recognition application is running at 22fps. The chip achieves up to 4.3× higher GOPS/W and 3.5× lower J/pixel compared with other parallel processors [2,4,6]. The low-latency NoC provides 76.8GB/s aggregated bandwidth and supports mesochronous communication without global clock synchronization. The NoC consumes 9% of the die area and 8% of the power consumption. A die photo and summary are shown in Fig. 16.2.7.

References:

[1] M.I. Posner and S.E. Petersen, "The Attention System in Human Brain," *Annual Review of Neuroscience*, vol. 13, pp. 25-42, 1990.

[2] 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, Feb. 2007.

[3] Kangmin Lee, et al., "Low-Power Networks-on-Chip for High-Performance SoC Design," *IEEE Trans. VLSI Systems*, vol. 14, pp. 148-160, Feb. 2006.

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

[5] A. Rodriguez-Vazquez, et al., "ACE16k: The Third Generation of Mixed-Signal SIMD-CNN ACE Chips Toward VSoCs," *IEEE Trans. Circuits and Systems-I: Fundamental Theory and Appl.*, vol. 51, pp. 851-863, May 2004.

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

978-1-4244-2010-0/08/$25.00 ©2008 IEEE

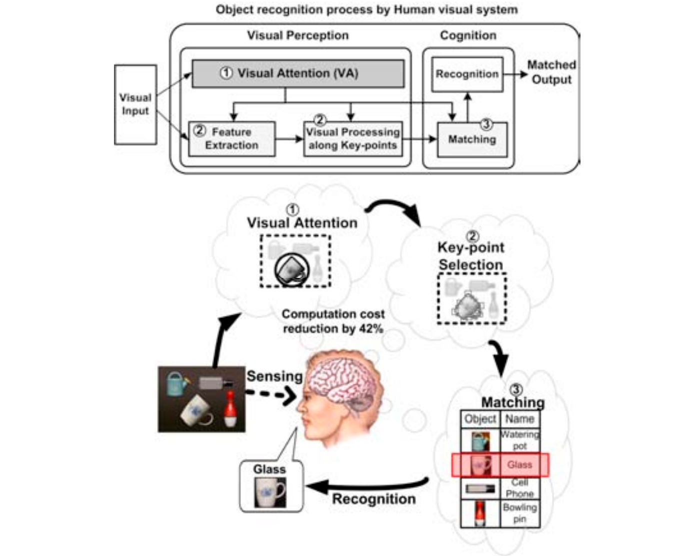


Figure 16.2.1: Object recognition system based on bio-inspired visual attention.

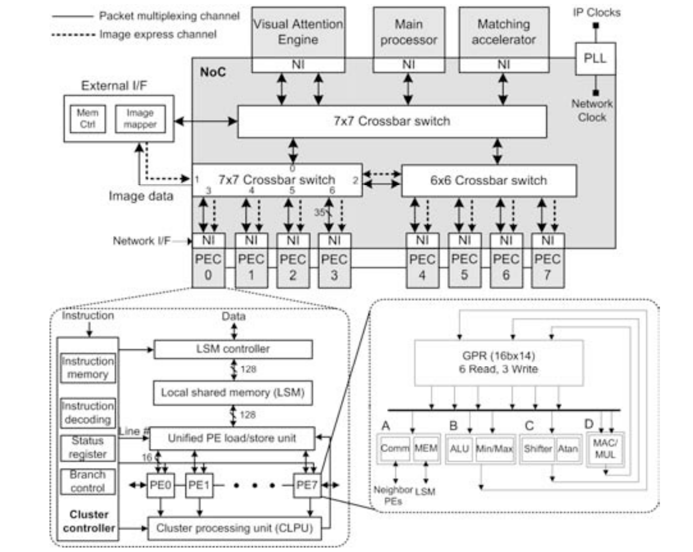


Figure 16.2.2: Block diagram of the overall system and PE clusters.

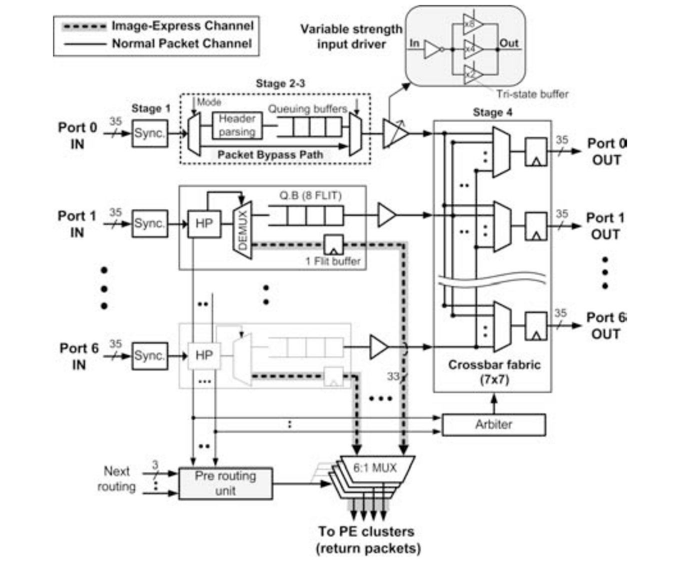


Figure 16.2.3: Proposed crossbar switch with dual channel and adaptive switching.

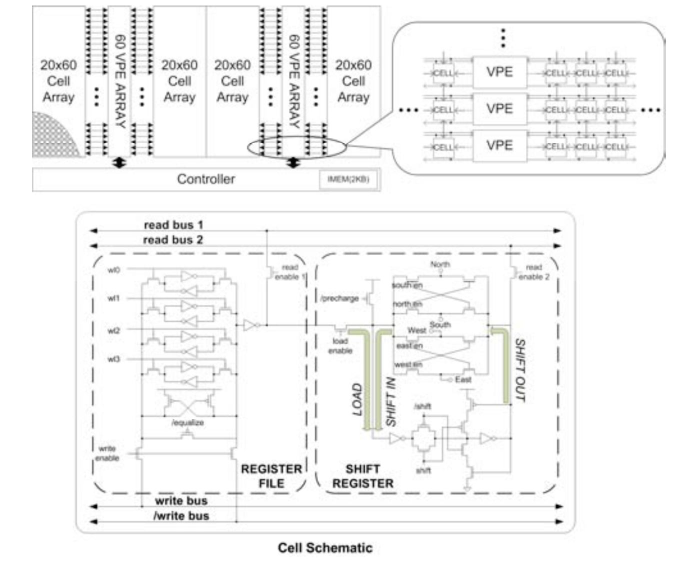


Figure 16.2.4: Circuit schematics of the VAE and its unit cell.

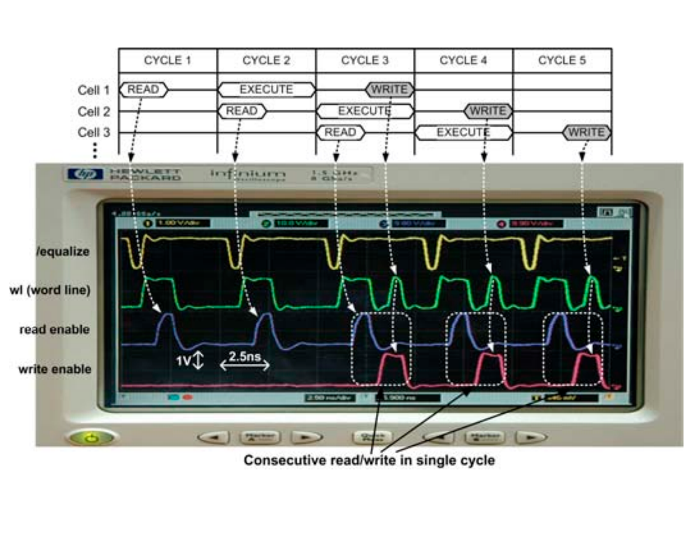


Figure 16.2.5: Measured waveforms of 3-stage pipelined operation of the VAE.

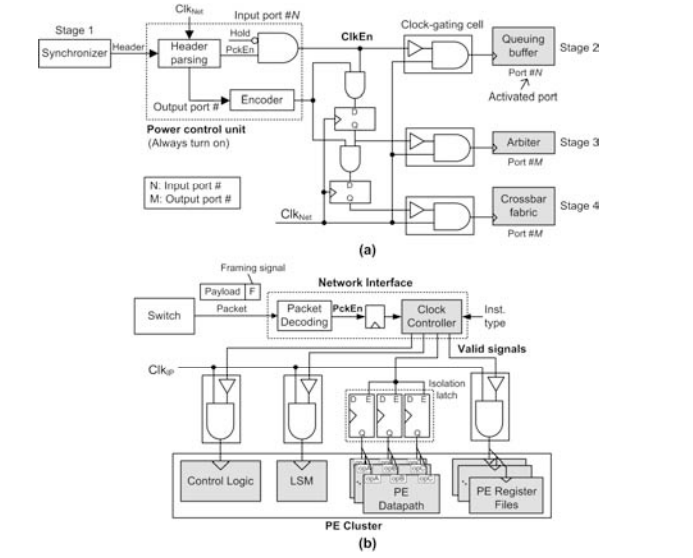
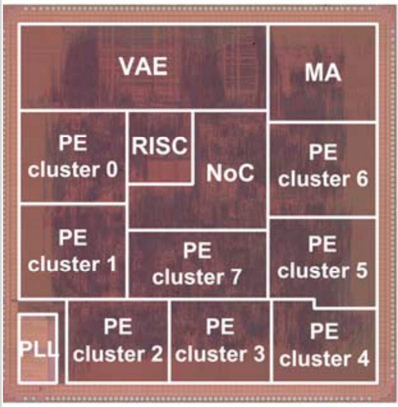
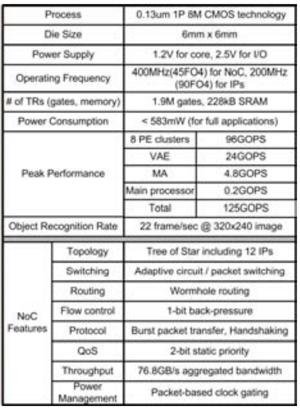


Figure 16.2.6: (a) Packet-based fine-grained clock gating and (b) IP-level power control.

Continued on Page 615





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(45FO4) for NoC, 200MHz (90FO4) for IPs	
# of TRs (gates, memory)	1.9M gates, 228kB SRAM	
Power Consumption	< 583mW (for full applications)	
Peak Performance	8 PE clusters	96GOPS
	VAE	24GOPS
	MA	4.8GOPS
	Main processor	0.2GOPS
	Total	125GOPS
Object Recognition Rate	22 frame/sec @ 320x240 image	
NoC Features	Topology	Tree of Star including 12 IPs
	Switching	Adaptive circuit / packet switching
	Routing	Wormhole routing
	Flow control	1-bit back-pressure
	Protocol	Burst packet transfer, Handshaking
	QoS	2-bit static priority
	Throughput	76.8GB/s aggregated bandwidth
	Power Management	Packet-based clock gating

Figure 16.2.7: Chip micrograph and summary.