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

Kwanho Kim, Seungjin Lee, Joo-Young Kim, Munsu 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 PEs: 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 NoC protocol supports burst packet transactions of up to 8 packets and handshaking for reliable transmission. A 7x7 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 pipeline stages of the switching 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 20x60 cells, 120 visual PEs (VPEs) shared between the cell arrays, and a controller. An 80x60 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 ST 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 neighboring cells, reducing the wiring complexity. As a result, the full-custom designed cell occupies a compact area of 496μm², 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 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. 16.2.6. 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 36mm² including 1.9M gates and 228kB SRAM. Power consumption is 583mW at 200MHz and 400MHz power supply (25˚C), while object recognition application is running at 220fps. The chip achieves up to 4.3x higher GOPS/W and 3.5x 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 8% of the die area and 8% of the power consumption. A die photo and summary are shown in Fig. 16.2.7.

References:
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
Figure 16.2.7: Chip micrograph and summary.