15.1 XETAL-II: A 107 GOPS, 600mW Massively-Parallel Processor for Video Scene Analysis

A. Abbo[1], R. Kleihorst[2], V. Choudhary[1], L. Sevat[2], P. Wielage[3], S. Mouy[3], M. Heijligers[3]

[1]Philips Research Laboratories, Eindhoven, The Netherlands
[2]NXP Corporate Research, Eindhoven, The Netherlands
[3]NXP Semiconductors, Eindhoven, The Netherlands

Advanced video processing applications such as object-based video coding and scene monitoring require extensive computational power to deal with the increasing complexity of the algorithms and the high pixel rates. The processing involves real-time robust detection of objects and their orientations in uncontrolled lighting conditions. Modern algorithms address this by performing multiple passes over the same frame at different scales and rotations using large sets of filters, e.g., Haar filters [1] for faces and blob filters for objects [2]. A useful property of these video analysis algorithms is the high degree of data-level parallelism present during the filtering and analysis stages. This parallelism has been exploited in building a power-efficient, fully programmable IC called Xetal-II, which provides a measured peak performance of 107 GOPS when operating at 84MHz and 1.2V, while dissipating less than 600mW. The IC is realized in 90nm CMOS and occupies $74mm^2$.

Decreasing CMOS feature size has become the enabler for low-cost and high-performance single-chip compute engines [3,4,6]. The limitations of [4] lie in the absence of a wide-word multiplier (due to the bit-serial architecture) and the 1Mb on-chip data, both of which reduce the overall performance when data-intensive video analysis applications are considered. While the architecture of [6] shows resemblance to the present design, there are significant differences with regard to the data path resolution, the number of on-chip compute units, the micro-architecture, the size of on-chip memory and the method of external stream interfacing.

Figure 15.1.1 shows the Xetal-II architecture, which is based on a massively-parallel single-instruction multiple-data (MP-SIMD) processing paradigm. A linear processing array (LPA) consisting of 320 processing elements (PEs) and 10Mb on-chip frame memory handles the compute-intensive data processing. The data input and output processors (DIP/DOP) provide an interface to 3 independent 10b video channels, each carrying up to 80Mpixels/s. The serial-to-parallel and parallel-to-serial stream conversion is done via two 320-pixel wide sequential line memories. A global control processor (GCP) manages operation of the IC. The chip is programmable using a sub-set of C, extended with a vector data type.

A summary of the chip characteristics is given in Fig. 15.1.2. The peak performance of 107 GOPS is achieved when the PEs execute instructions with 4 operations at 84MHz clock frequency. At this frequency, applying a non-separable 11×11 filter kernel on a VGA frame takes 2ms to complete, allowing 500fps at 100% processor utilization. A 3×3 median filter on 1280×720 HDTV frame takes 4ms per colour component when executing at full speed.

Although the MP-SIMD paradigm simplifies the processor design due to the reduced operating frequency, it introduces interconnect complexity due to the massively parallel memory access. At 84MHz, the Xetal-II memory interface provides 1.3Tb/s over 15360 wires. To deal with this complexity in a power-efficient way, the design has been partitioned into 40 compute tiles each with 8 PEs and associated data memory, as shown in Fig. 15.1.3. The data memory has a pseudo-dual port interface to provide read and write accesses in one cycle. Neighbouring PEs can read input data of one another, which is also extended as inter-tile communication. Each tile performs flag addition and flag-based result selection in which the result of the right-most PE whose flag is set is selected. The flag sum and the selected result are sent to the GCP to facilitate object or event analysis. To allow performance adaptation, the tile is divided into logic and memory voltage domains, coupled with level-shifters.

The PE, shown in Fig. 15.1.4, is a 16b compute unit with a local register for immediate result feedback and a flag for guarded executions. The PE handles 64 instructions ranging from logical operations and multiply-accumulate (MAC) to compound instructions (up to 4 operations), all executed in one cycle.

The main tasks of the GCP, shown in Fig. 15.1.5, include controlling the program flow, handling interrupts from DIP and DOP and signal-processing tasks like analyzing the result of the LPA and updating filter coefficients. The GCP is a 16b processor with 36 instructions that operates on a data memory of 2048 words accessed via 3 read and 1 write buses. Base plus indexed address computation is possible using registers in the Special-Purpose Register File (SPRF). The instruction fetch and decode units allow parallel execution of instructions on the GCP and the LPA and support two threads of execution to handle high-priority (time-critical) and low-priority tasks. Separate program counters and SPRF blocks are provided to enable task switching with zero time-overhead.

Traditionally, it is very expensive to give Look-Up-Table (LUT) functionality to an SIMD processor since every PE needs its own table and an extra indirect address path. In Xetal-II, the LUT functionality is performed serially using a shared table and an internal feedback path that connects the DOP and DIP units as shown in Fig. 15.1.1. The LUT, located in the DOP unit, is configurable by the GCP and the conversion takes place in the background while the LPA processes other data.

Figure 15.1.6 shows the power management topology and the measured performance of the chip when executing a 16b MAC operation. When full performance is not demanded, adapting the operating voltage and frequency would improve the performance efficiency [5]. The core of the chip has 2 voltage domains: V_{DVS}(0.7 – 1.2V) for the logic in the LPA and V_{MEM}(0.9 – 1.2V) for the memories and the remaining logic. The measured speed of 84MHz at 1.2V is achieved with half the power pads unconnected since a 156-pin package was chosen to be pin-compatible with previous designs. When all the power pads are connected, the chip is expected to operate at 150MHz.

Figure 15.1.7 shows the die micrograph in which the tile-based layout of the chip is highlighted. The GCP is placed at the center to keep the length of interconnect to the PEs short and facilitate clock skew balancing.

With the achievements described here and the advantages of modern silicon technology, advanced video processing tasks, such as real-time full-frame video analysis, become feasible for portable consumer applications.

References:
[1] M. Jones, P. Viola, "Fast Multi-View Face Detection," *Mistubishi Electronics Research Laboratory, TR2003-96,* June 2003.
[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. of Computer Vision*, vol. 60, no. 2, pp. 91-110, Nov., 2004.
[3] R. P. Kleihorst, et al., "XETAL, a Low-Power High-Performance Smart Camera Processor," *Proc. ISCAS,* pp. 215-218, May, 2001.
[4] M. Nakajima, et al., "A 40GOPS 250mW Massively Parallel Processor Based on Matrix Architecture," *ISSCC Dig. Tech. Papers,* pp. 410-411, Feb., 2006.
[5] A. A. Abbo, et. al., "Power Consumption of Performance-Scaled SIMD Processors," *Proc. of PATMOS,* pp. 532-540, Sep., 2004.
[6] Shorin Kyo, et. al., "An Integrated Memory Array Processor for Embedded Image Recognition Systems," *Proc. ISCA,* pp. 134-145, June, 2005.

 1-4244-0852-0/07/$25.00

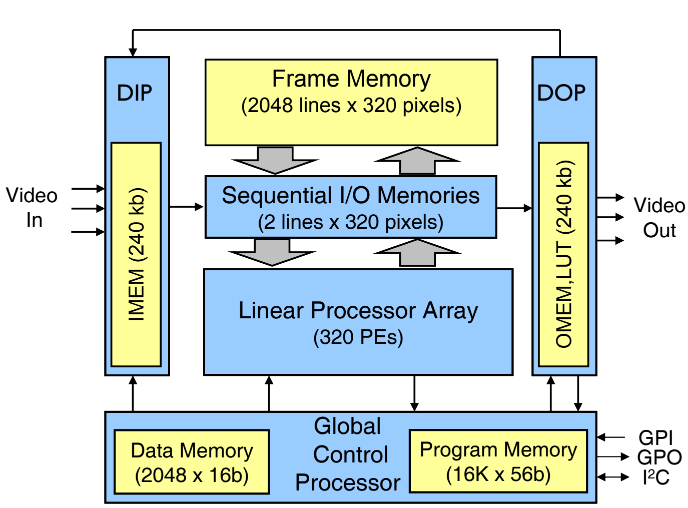

Figure 15.1.1: Block diagram of the Xetal-II architecture.

Technology	90nm CMOS
Chip Area	74mm²
Number of Gates	740 kgates
On-Chip Data Memory	10Mb (4 VGA frames)
On-Chip Program Memory	16384 Instructions
Measured Clock Speed	84MHz
Measured Peak Performance	107 GOPS
Internal Data Bandwidth	1.3Tb/s
Data-path Resolution	16b
Input Video Port	3x10b
Output Video/Data Port	3x10b
Number of Clock Domains	9
Number of Voltage Domains	2
Measured Power Consumption	600mW

Figure 15.1.2: Chip features.

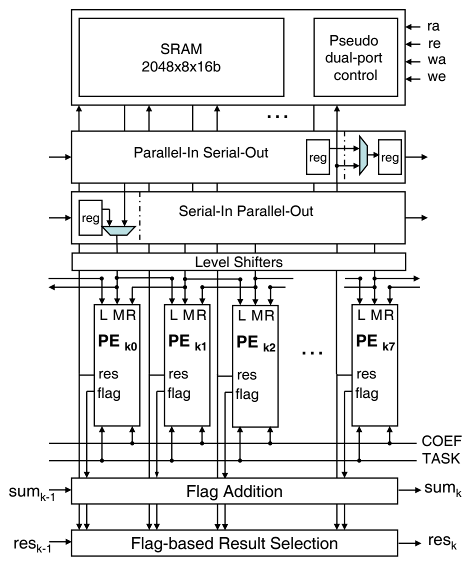

Figure 15.1.3: Architecture of a Xetal-II compute tile.

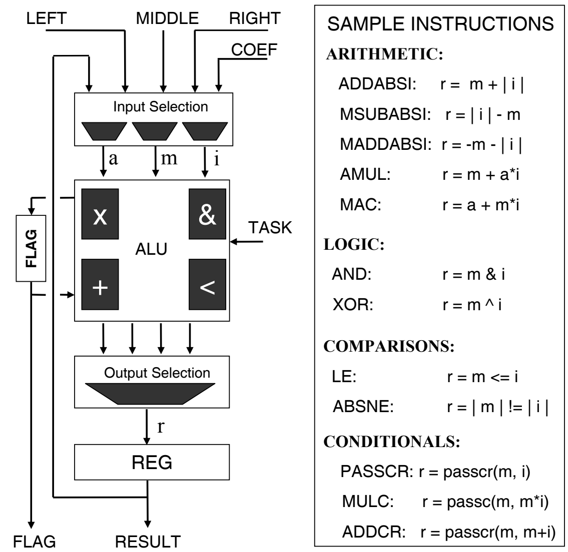

Figure 15.1.4: Processing Element (PE) architecture.

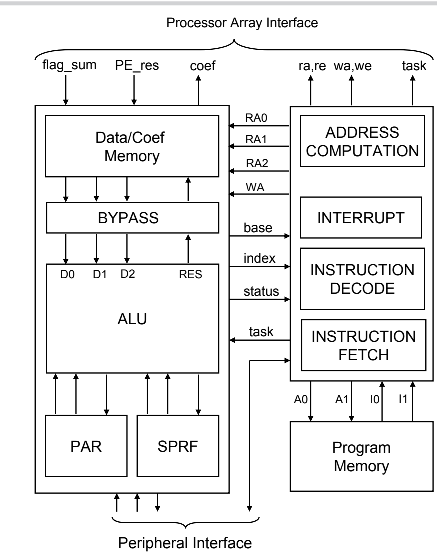

Figure 15.1.5: Global Control Processor (GCP) architecture.

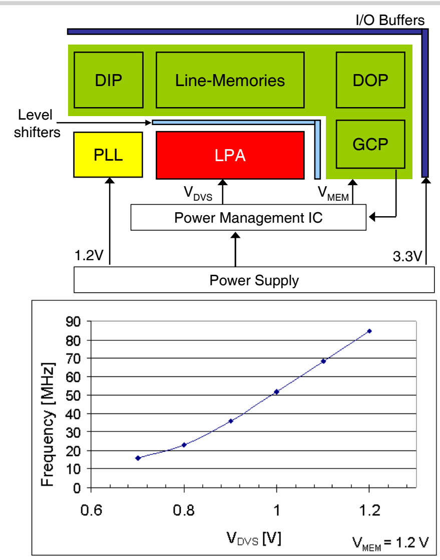

Figure 15.1.6: Power management and measured performance.

Continued on Page 602

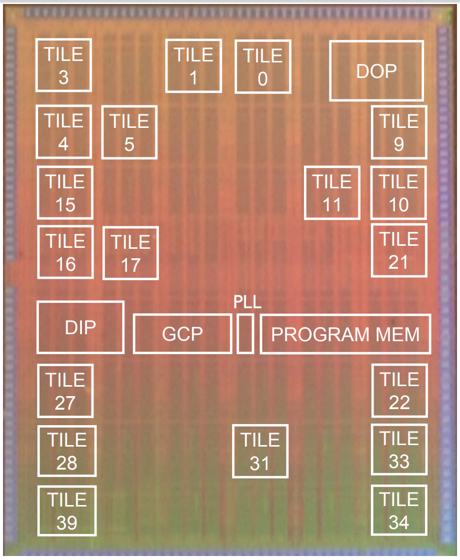

Figure 15.1.7: Die micrograph.

