# Architecture and Design of a Programmable 3D-Integrated Cellular Processor Array for Image Processing

Alexey Lopich Piotr Dudek
School of Electrical and Electronic Engineering
The University of Manchester
United Kingdom
{a.lopich, p.dudek}@manchester.ac.uk

*Abstract*— **In this work we present a design of a massively-parallel cellular processor array implemented in 3D CMOS technology. The proof of concept 128×96 array device is partitioned across two custom designed layers. Additionally, three layers of DDR memory are vertically stacked and bonded underneath. The processor benefits from 358Gbit/s data rate between memory and array, as well as from high logic density, thanks to improved routing across silicon layers with Trough Silicon Vias (TSVs).**

## I. INTRODUCTION

The merits of massively parallel computing have been explored for several decades [1, 2]. In particular, the advantages have been achieved in low-level image processing, where regularity of image data and full-scale parallelism of processing algorithms naturally maps onto fine-grain processor-per-pixel architectures. Apart from significant performance gains, the benefit of this approach is the elimination of data-transfer bottleneck thus increasing systems throughput. The architecture of such devices is typically based on a SIMD cellular processor arrays (CPA), with either linear processor-per-column architecture for line-by-line processing [3, 4] or fine-grain processor-per-pixel implementation [5, 6]. Many SIMD processor arrays have been designed as stand-alone image processing devises. Examples of such devices can be found in [3, 5-7]. Further development has led to integration of image sensing and parallel in-pixel processing into a single silicon device, often referred to as a vision chip, a smart sensor or a focal-plane processor [8, 9]. Such device, after receiving and parallel-processing of large amounts of image data can output only scalar descriptors (e.g. target coordinates, histogram, object size, etc.) [10, 11] for further processing by serial architectures. In the past we have developed and implemented several pixel-parallel architectures, where sensor is integrated with simple processing element (PE) [11, 12]. While featuring exceptional power efficiency and area utilisation, the main disadvantage of such approach is significant sacrifice of optical properties of the device namely fill-factor, due to the location of processing circuit adjacent to sensor and sensitivity due to utilisation of standard (not CMOS Image Sensor) technology required to implement the processing part. Finally the actual pixel pitch is constrained by required in-pixel functionality, making fabrication of vision chips with large resolution not economically viable.

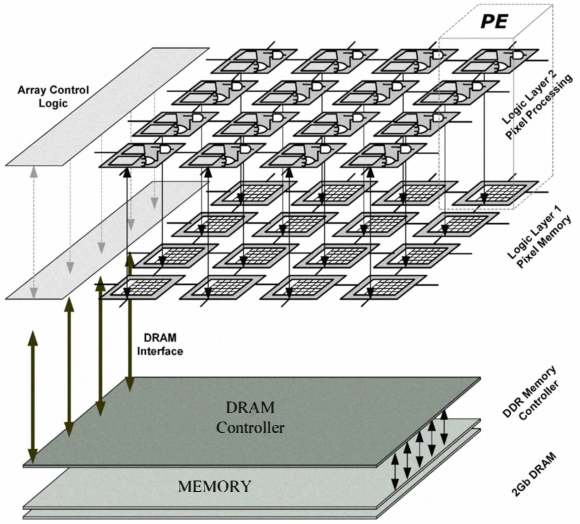


Figure 1. Architecture of the proposed cellular processor array. Processing cell is distributed across 2 layers.

Aforementioned constrains, however, are primarily imposed by the planarity of the fabrication technology. Current advances and availability of 3D silicon integration with randomly placed TSVs with relatively small pitch offer a natural solution to the above mentioned issues. By appropriate partitioning of processing cell's architecture across silicon layers (potentially fabricated on various processes technologies) it is possible to design compact processing cells suitable for building high-resolution smart sensors [13]. However, the advantages of 3D technology can be applied for building non-optical devices as well, as it brings opportunities for revolutionary designs, both on the microarchitecture level (distributing the processor cells across many layers, alleviating the constraint on the complexity and local memory capacity in the PE), and on the system level (integrating hierarchy of processors, from fine-grain to coarse-grain arrays, together with controllers and on-chip memory). In particular, implementation of a fine-grain CPA in a 3D technology offers unique opportunity to exploit the potential processor-memory bandwidth and to optimise communications, providing alternative but potentially far more powerful architecture than

978-1-4577-0170-2/11/$26.00 ©2011 IEEE

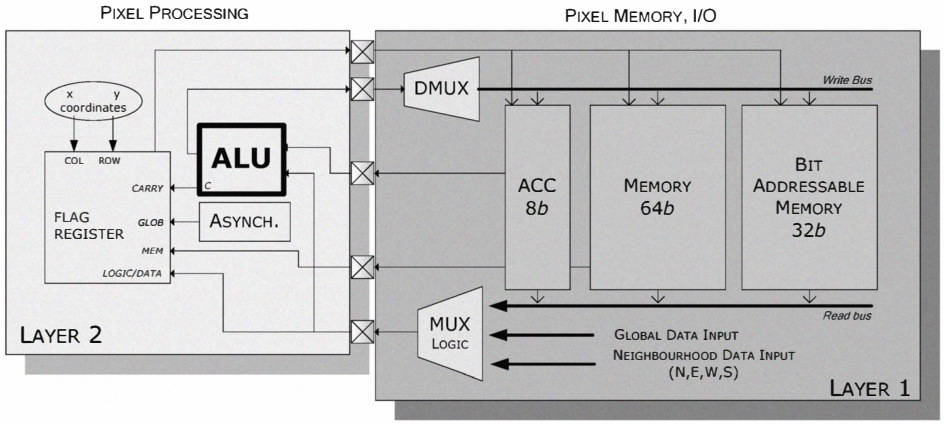


Figure 2: The block diagram of processing element. The design is partition circuitry is physically placed on Layer 1; bit-serial units, i.e. ALU, flag register, asynchronous processors are placed on Layer 2.

the relatively coarse-grain, many-core parallel approaches currently considered by industry.

In this work we present architecture of cellular processor array implemented on two layers of stacked silicon. Although overall architecture is partly inspired by our previous research [11, 12], significant improvements have been made thanks to an adaptation onto novel 3D IC fabrication process. The design has been implemented in low-power 0.13 µm CMOS process and followed by subsequent face-to-face bonding using state-of-the-art processes provided by Tezzaron Semiconductor [14]. The functionality of the chip is significantly expanded by the availability of additional 2Gbit of dynamic RAM, vertically stacked and connected via TSVs to the processor array.

## II. SYSTEM ARCHITECTURE

The outline of the proposed architecture is depicted in Figure 1. The image data is uploaded from external source outside the chip. The array operates in SIMD mode, where each processing node executes identical program broadcast by a central controller. Each processing cell has individual flag indicator, which enables control over executed operations, providing some degree of local autonomy. The PE operates as a simple digital microprocessor and comprises register memory bank, ALU, Flag Register, local and global data I/O circuitry and auxiliary unit for global asynchronous processing. Local memory is organised in thirteen blocks of 8-bits and can be accessed either in bit-parallel manner or by individual bits. This approach enables fast data I/O operations together with compact design of bit-serial ALU. Local/neighbour/global data manipulations consist of all arithmetic and logic functions, of which the majority is executed in two clock cycles (per bit). In addition to pixel processing, the presented array features random pixel access, flexible pixel/block addressing mechanism [15] and asynchronous address extraction facilities, thus enabling address event representation (AER) readout scenario [16].

## III. DESIGN FEATURES

Since current prototype implementation does not feature in-pixel sensors, the interface between the chip and an external device (controller with a stand-alone CMOS sensor) is optimised to provide high data transfer rates. The on-chip controller facilitates a 32-bit interface to an external device. Single 32-bit wide data transaction requires one clock cycle (the nominal clock frequency is 350 MHz). The addressing mechanism allows addressing four pixels at a time, so that 32-bit input data, which is transmitted to four neighbouring rows (<0-3>, <4-7>,…,<124-127>) is written to four cells at once, thus providing 5.5 Gbit/s or $6\times10^4$ fps. The 32-bit interface is chosen for practical reasons, to minimise external connectivity, however a wider data bus (up to 1024 bits) could be easily implemented for higher throughput. In future implementations

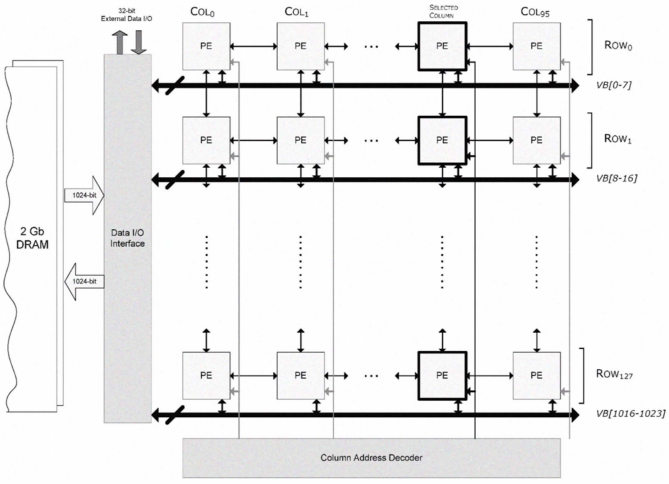


Figure 3: Schematic representation of data I/O architecitre

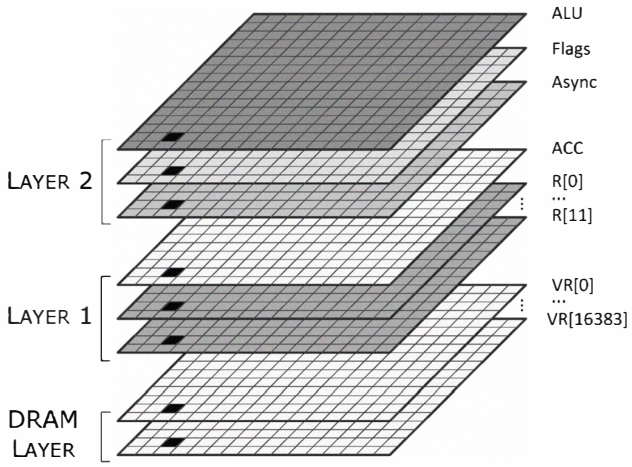


Figure 4: Register plane of the cellular array architecutre. A single processing element is marked.

we envisage that an image sensor with suitable digital/analogue interface could be directly bonded on top of the processor array, with either per-pixel or column-based connectivity to the processor array.

The chip benefits from 3 additional DRAM layers, vertically bonded to the main die. These layers contain 2Gbit of memory together with memory controllers. The memory has a 1024-bit interface, so that an entire column of data (128 8-bit values) can be written/read in a single transaction in two clock cycles (Figure 3). The additional DRAM memory is treated as a set of "virtual" register planes and acts as a memory extension for processing cells. Thus, in addition to thirteen 8-bit registers every PE has 16384 virtual registers (Figure 4), which are located in DRAM and can be accessed at random times. In order to enable efficient use of virtual memory, the interface between the array and memory is optimised, so that access to a virtual register, which implies reading out/in the entire 128×96 register plane, takes less than 300 ns. The use of virtual registers removes the common local memory limitation of pixel-parallel CPA's, which have to trade-off local memory capacity and physical cell area.

The bandwidth of the interface between cellular array and memory is 358 Gbit/sec. The column selection is performed by an address mechanism described in our prior work [15]. The address is supplied by external controller, through an 8-bit address port.

In order to reduce the timing overhead for data swap operations between register and DRAM, all I/O routines can be conducted in parallel with processing. It has been achieved, by dedicating a separate register with independent select, read, write signals as well as separate data input/output. The register is connected to a virtual bus (VB) as well as internal data buses. If the operation requires data stored in several virtual registers, these registers have to be fetched and buffered inside the pixel.

## IV. IMPLEMENTATION

The proof-of-concept design, which incorporates 128×96 processing elements, has been implemented on two vertically integrated 5×5 $mm^2$ silicon dies. It utilises 130 nm low-power CMOS process (supply voltage 1.5V) and face-to-face and TSV stacking techniques provided by Tezzaron [14]. Every PE is spanned across two layers and has an area of 30×30 $\mu m^2$ on each layer (Figure 5). The inter-tier interconnection is implemented by connecting to the top metal contacts on each logic die, which are subsequently aligned and stacked face-to-face to form the link. The pitch of this interconnection is 2.4 µm. The stacked devices are then diced and stacked onto the DRAM memory layer in a die-to-wafer process. The latter is based on connecting the Metal 1 to the Backside Metal on the Layer 1 via TSVs and subsequent back-to-face bump-bonding (with the pitch of microbumps 25 µm) to the top metal of the wafer, which contains DRAM memory. The memory wafer is larger than the designed chip and also contains a top-metal pattern for external data I/O. These metal pads are routed to connect via TSVs to the I/O cells placed on Layer 1.

While the size of inter-tier contact pitch is continuously

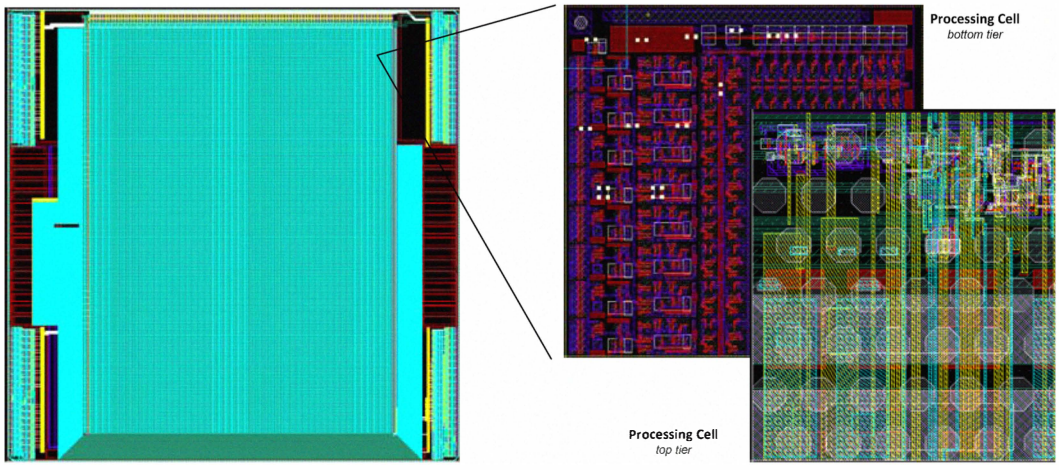


Figure 5: The layout of the chip and processing cell logic, distributed across two physical layers. Octagonal bondpads on top metal, displayed on top tier, are used for fase-to-face bonding to bottom tier. The latter has similar top metal pattern, however it is not displayed for clarity.

decreasing as 3D stacking technology progresses (for current process it is 2.4×2.4 $\mu m^2$), it still relatively large compared to metal interconnect vias on standard CMOS fabrication process (~0.3 $\mu m^2$). From the compact processing pixel cell design perspective it imposes certain constraints on partitioning the design in order to minimise the number of inter-tier interconnections. To achieve this goal the functionality of the cell is split in such a way that communication between tiers is performed in a bit serial manner. Layer 1 is dedicated to local pixel memory, controller for local data buses and data I/O circuitry, while Layer 2 accommodates ALU, Flag Register and auxiliary unit for global asynchronous data processing. Such division also simplifies physical custom design and minimises the overhead associated with global control signal routing. It can be noted that local PE's memory has a compact layout implementation with dense yet regular control/address signal routing. At the same time, Layer 2 represents modular design, where each block has its own control signals and requires individual placement. Thanks to such partitioning, each cell incorporates 568 transistors on Layer 1 (memory) and 98 transistors on Layer 2 (although not entire 30×30 $\mu m^2$ area is occupied).

The design has been taped-out, and is currently in the process of 3D stacking. Post-layout simulations indicate correct operation at 350 MHz clock, thus delivering 274 GOPS for greyscale operations and more that 2 TOPS for binary processing. Global asynchronous processing capabilities provide further performance gains for global image processing algorithms [17]. The performance on common operations is summarised in Table I.

TABLE I: Expected performance of the processor array in most common operations. Operation at 350MHz clock is assumed.

| **Operation** | **Time (per frame)** | **Comment** |
|---|---|---|
| **Logic/Conditional** | 5.8 ns | 2 clock cycles |
| **Addition/Subtraction** | 49.6 ns | 8-bit |
| **Multiplication** | 443.2 µs | Unsigned |
| **Global Sum** | 1.77 µs | worst case 22-bit |
| **Pixel Count** | 1.12 µs | 14-bit |
| **Coordinate Extraction** | 214 ns | Four border coordinates for binary object |
| **Asynchronous trigger-wave propagation** | 100.8 ns | ~0.45 ns delay per cell |
| **I/O Stream** | | |
| **DRAM** | 22.8 GPixel/s | 8-bit greyscale |
| **External** | 350 MPixel/sec | 8-bit greyscale |

## V. Conclusions

We have implemented an SIMD cellular processor array in 3D silicon integration technology, which provided the possibility to partition the design, so that each part can have the most compact physical implementation. The advantage of utilizing vertical integration technology and more specifically small pitch of inter-tier interconnection have been demonstrated by achieving small processing cell area, while significantly expanding overall functionality and performance compared to previous planar realizations. In addition to two custom design logic layers, the chip benefits from additional three layers of vertically integrated DRAM memory, which further expands potential applications requiring processing of large amounts of temporary data, such as complex intra-frame video analysis. From this work, we can observe that pixel-parallel cellular processor arrays are naturally suitable for 3D integration and can be easily scalable to more layers, especially when various process technologies can be integrated together.

## VI. Acknowledgment

This research was sponsored by EPSRC grant no. EP/H017453/1.

## References

[1] K.E. Batcher, *Design of a massively parallel processor.* IEEE Transactions on Computers, 1980. **c-29**(9): p. 836-840.

[2] M.J.B. Duff and D.M. Watson, *The cellular logic array image processor.* Computer Journal, 1977. **20**(1): p. 68-72.

[3] A.A. Abbo, R.P. Kleihorst, V. Choudhary, L. Sevat, P. Wielage, S. Mouy, et al., *Xetal-ii: A 107 gops, 600 mw massively parallel processor for video scene analysis.* IEEE Journal of Solid-State Circuits, 2008. **43**(1): p. 192-201.

[4] L. Lindgren, J. Melander, R. Johansson, and B. Moller, *A multiresolution 100-gops 4-gpixels/s programmable smart vision sensor for multisense imaging.* IEEE Journal of Solid-State Circuits, 2005. **40**(6): p. 1350-1359.

[5] D. Andrews, C. Kancler, and B. Wealand. *An embedded real-time simd processor array for image processing*. in Proceedings of the 4th International Workshop on Parallel and Distributed Real-Time Systems. 1996. Los Alamitos, USA: p 131-134.

[6] J.C. Gealow and C.G. Sodini, *Pixel-parallel image processor using logic pitch-matched to dynamic memory.* IEEE Journal of Solid-State Circuits, 1999. **34**(6): p. 831-839.

[7] R.E. Morley, Jr. and T.J. Sullivan. *A massively parallel systolic array processor system*. in *Proceedings of the International Conference on Systolic Arrays,* 1988: p 217-225.

[8] Á. Zarándy, *Focal-plane sensor-processor chips*. 1st ed. 2011: Springer. 305.

[9] A. Moini, *Vision chips or seeing silicon*. 1999: Kluwer Academic Publishers

[10] T. Komuro, I. Ishii, M. Ishikawa, and A. Yoshida, *A digital vision chip specialized for high-speed target tracking.* IEEE Transactions on Electron Devices, 2003. **50**(1): p. 191-199.

[11] A. Lopich and P. Dudek, *A SIMD cellular processor array vision chip with asynchronous processing capabilities.* IEEE Transactions

on Circuits and Systems I: Regular Papers, 2011(99): p. DOI 10.1109/TCSI.2011.2131370

[12] P. Dudek and S.J. Carey, *General-purpose 128128 simd processor array with integrated image sensor*. Electronics Letters, 2006. **42**(12): p. 678-679.

[13] P. Dudek, A. Lopich, and V. Gruev. *A pixel-parallel cellular processor array in a stacked three-layer 3D silicon-on-insulator technology*. in European Conference on Circuit Theory and Design. 2009: p 193-196.

[14] S. Gupta, M. Hilbert, S. Hong, and R. Patti, *Techniques for producing 3D ICs with high-density interconnect*, in International VLSI Multilevel Interconnection Conference. 2004: Waikoloa Beach, HI, USA.

[15] P. Dudek. *A flexible global readout architecture for an analogue simd vision chip*. in IEEE International Symposium on Circuits and Systems, . 2003. Bangkok, Thailand: p 782-785.

[16] A. Lopich and P. Dudek. *An 80x80 general-purpose digital vision chip in 0.18 um cmos technology*. in IEEE International Symposium on Circuits and Systems. 2010. Paris, France: p 4257-4260.

[17] A. Lopich and P. Dudek, *Global operations in simd cellular processor arrays employing functional asynchronism*, in IEEE International Workshop on Computer Architecture for Machine Perception and Sensing. 2007: Montreal, Canada. p. 16-23.