7.3 A 275mW Heterogeneous Multimedia Processor for IC-Stacking on Si-Interposer

Hyo-Eun Kim, Jae-Sung Yoon, Kyu-Dong Hwang, Young-Jun Kim, Jun-Seok Park, Lee-Sup Kim

KAIST, Daejeon, Korea

Most data-intensive operations for multimedia applications such as image processing, vision, and 3D graphics require high external memory bandwidth. In augmented-reality (AR) processors [1], both 3D graphics and vision operations are required, so memory bandwidth becomes even more critical. In [1], however, memory bandwidth is not considered, floating-point processing is not supported, and there is no cache memory for texturing, which is a performance bottleneck of common graphics pipelines. In this work, a heterogeneous multimedia processor is presented to process various mobile multimedia applications in a single chip on Si-interposer for high memory bandwidth. The implemented processor has 4 key features: (1) A transceiver pool (TRx) that reconfigures strength of output drivers according to the channel loss for IC-stacking on Si-interposer, (2) A mode-configurable vector processing unit (MCVPU) for frame-level parallelism, (3) An energy-efficient unified filtering unit (UFU) with adaptive block selection (ABS) algorithm for memory-access-efficient texturing, and (4) a unified shader (US) with floating-point scalar processing elements (SPE) and partial special function units (PSFU) to enhance graphics processing performance and quality. With these techniques, we achieve 1.7× frame rate and 8× memory bandwidth improvement in full AR operation.

Figure 7.3.1 shows a target prototype of the entire system. The processor and DRAM silicon dies are stacked on a Si-interposer redistribution-layer. A TRx based on channel characteristics of the Si-interposer enables data communication between two dies while minimizing channel loss. Through-silicon-vias (TSV) constitute the interface for external data communication and power supply. The processor is composed of 4 programmable IPs (MCVPUs, a UFU, a US, and a RISC), 3 application-specific IPs (vision/graphics specific processing units (VSPU/GSPU), and a pose estimator), and a custom-designed analog IP, TRx. MCVPUs perform both data-intensive low-level image processing and per-pixel operations for vision and 3D graphics, respectively. Sixteen types of filtering operations for vision and 3D graphics are supported in UFU with a level-0 data buffer for energy-efficiency. The US performs floating-point geometry processing. The VSPU receives parallel data from MCVPUs and generates descriptors for pose estimation. The pose estimator makes a transformation matrix using descriptor vectors, and stores it into the constant-memory of the US for matrix multiplication. The GSPU interpolates vertex attributes for fragment rasterization and generates each fragment's level-of-detail (LOD) of texture maps for the ABS algorithm. The ABS algorithm includes tile-based rasterization (TBR), which exploits spatial locality of pixel fragments [3]. Each IP is independently controlled by IP-level clock gating, so different operation modes can be organized with different combination of various IPs.

In general, a processor communicates with an external memory through metal wires on a PCB. The length of metal wire is normally centimeter-order, so the channel loss is considerable when using high-speed data communication. Stacking the processor and DRAM dies on Si-interposer overcomes this drawback, since the channel length of Si-interposer is much shorter than a PCB's. It also realizes system miniaturization [2]. TRx is designed for channel (1.6GHz) loss compensation of Si-interposer as shown in Fig. 7.3.2. Eight select-signals are generated from a linear-feedback-shift-register, and serialize multiple data bits into a single output stream. Strength of output drivers is configured as 8 steps according to the channel loss. A serialized input stream is demuxed into parallel streams in a receiver. TRx results in 8× memory bandwidth improvement compared to previous work [1].

Descriptor generation with pose estimation and 3D object augmentation are completely independent operations, so these 2 operations can be processed in parallel. In this case, performance is substantially improved, but required memory bandwidth also increases linearly. So, our system is appropriate for frame-level parallelism. Four MCVPUs and a UFU can be configured as independent clusters for frame-level parallelism of AR. The MCVPU consists of 8 PEs, each of which has 6-way VLIW architecture for 16b arithmetic, logic operations, and data communication. Eight PEs are organized as an 8-way SIMD unit for low-level image processing (IMG-mode), and two 4-way SIMD units for 3D graphics (3D-mode). An image-processing operation configures 4 MCVPUs as IMG-mode, and uses UFU as a filtering unit for low-level image processing. A 3D graphics operation configures 4 MCVPUs as 3D-mode for pixel shading, and uses UFU as a texture unit. Three MCVPUs in IMG-mode and MCVPU3 with UFU in 3D-mode are organized as independent processing clusters for frame-level pipelining as shown in Fig. 7.3.1. Frame-level pipelining requires a large set of registers and memory spaces. This increases energy and area cost. In our processor, however, CMEM of US is used for frame-level pipelining. The proposed AR configuration achieves 1.7× higher frame rate compared to previous work [1].

Most low-level vision operations consist of various filtering operations. For a single filtering operation, multiple instructions are required in a general processing core. In UFU, however, a single instruction is enough for each filtering operation. Figure 7.3.3 shows the UFU. The UFU supports 16 filtering operations for graphics and vision. The previously proposed CFU supports 7 filtering operations [4]. UFU includes a 512B L0 buffer and an 8KB L1 cache. The L0 buffer improves energy efficiency with the addition of small hardware area by limiting direct references to L1 SRAM cache. In a texturing operation, energy consumed in the UFU is reduced by 80% compared to the conventional texture unit, and 15% compared to the TFM [5]. The ABS algorithm dynamically controls the size of a texture block to be fetched from an external memory. It reduces stall cycles caused by external memory accesses, and then enhances texturing performance. The center point of each tile, area and shape of the current triangle determine the size of a texture block to be fetched as described in Fig. 7.3.4. Texturing performance is improved by 17.9% compared to the A-index [6], and 9.1% compared to the TBR without ABS [3].

In a general graphics pipeline, a floating-point datapath is used in geometry processing for its precision. The US consists of 4 homogeneous single-precision floating-point SPEs, and each SPE has a 4-way VLIW architecture. Each SPE has its own PSFU, but the look-up table (LUT) is shared by 4 SPEs with table loader (TBLD) as shown in Fig. 7.3.5. TBLD exploits access patterns to the shared LUT by using multi-fetch and bank-partition schemes [7]. The US reduces total latency of full shading operation by 53%, 30% compared to 1-issue SIMD and 2-issue SIMD architecture, respectively.

Figure 7.3.6 summarizes chip features. The UFU with the ABS algorithm improves energy efficiency and performance, so the energy-delay product is reduced by 70%. The US reduces the area-delay product by 38.5% compared to the 2-issue VLIW SIMD. This chip is fabricated in 0.13μm 1P6M CMOS. It integrates 1.46M logic gates and an analog IP, TRx, within a 4×4mm^2 die. MCVPUs operate at maximum 200MHz, and the other digital IPs operate at 100MHz.

Acknowledgements:
Chip fabrication was supported by IDEC at Korea Advanced Institute of Science and Technology (KAIST), Korea. This work was supported by the IT R&D program of MKE/KEIT. [KI002134, Wafer Level 3D IC Design and Integration]

References:

[1] Jae-Sung Yoon, et al., "A Graphics and Vision Unified Processor with 0.89uW/fps Pose Estimation Engine for Augmented Reality," *ISSCC Dig. Tech. Papers*, pp. 336-337, Feb. 2010.

[2] J.U. Knickerbocker, et al., "3D Silicon Integration," *Proc. of Electronic Components and Technology Conf.*, pp. 538-543, May 2008.

[3] J. McCormack, et al., "Tiled polygon traversal using half-plane edge functions," *Proc. of SIGGRAPH Conf. on Graphics Hardware*, pp. 15-21, 2000.

[4] Chih-Hao Sun, et al., "CFU: Multi-Purpose Configurable Filtering Unit for Mobile Multimedia Applications on Graphics Hardware," *Proc. of Conf. on High Performance Graphics*, pp.29-36, 2009.

[5] B. V. N. Silpa, et al., "Texture Filter Memory – A power-efficient and scalable texture memory architecture for mobile graphics processors," *Proc. of Int. Conf. on Computer-Aided Design*, pp. 559-564, Nov. 2008.

[6] C. H. Kim, et al., "Adaptive selection of an index in a texture cache, " *Proc. of Int. Conf. on Computer Design*, pp. 295-300, Oct. 2004.

[7] Young-Jun Kim, et al., "Bank-Partition and Multi-Fetch Scheme for Floating-Point Special Function Units in Multi-Core Systems," *Proc. of Int. Symp. on Circuits and Systems*, pp. 1803-1806, May 2009.

 978-1-61284-302-5/11/$26.00

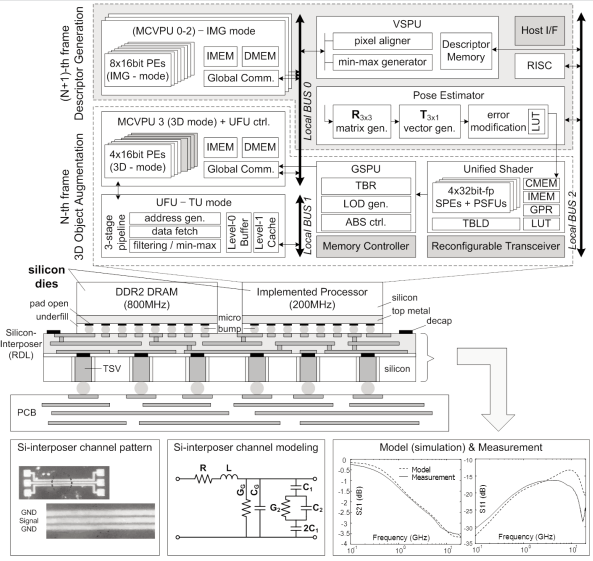

Figure 7.3.1: System prototype and architecture of the processor.

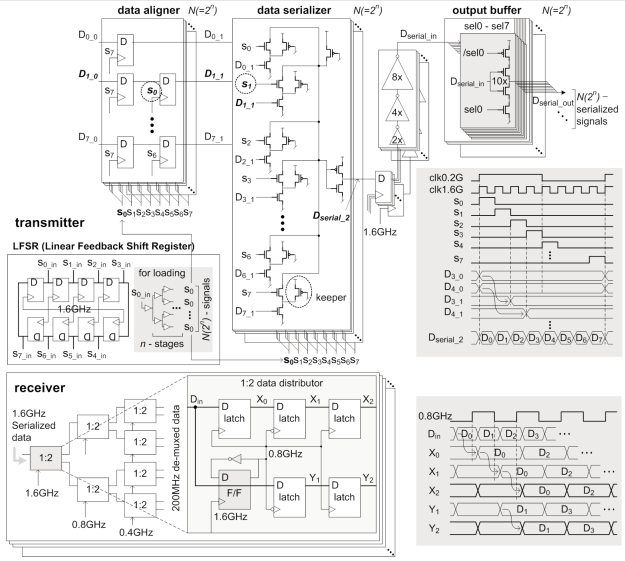

Figure 7.3.2: Transceiver with reconfigurable output driver.

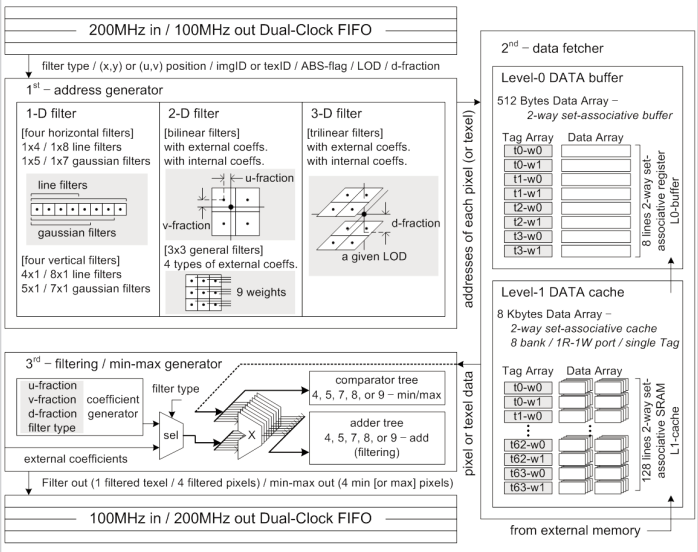

Figure 7.3.3: UFU for low-level filtering operations.

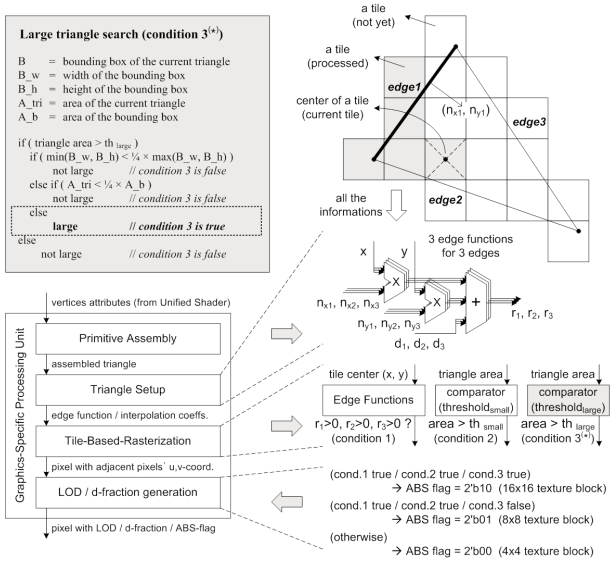

Figure 7.3.4: ABS algorithm for memory-access-efficient texturing.

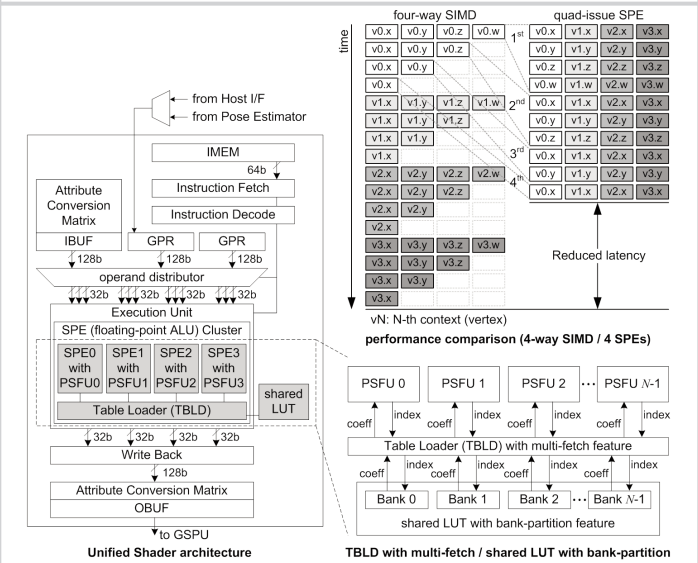

Figure 7.3.5: US architecture with PSFU.

Process Technology	0.13um CMOS (1P6M)
Power Supply	1.2V (core) / 3.3V (IO)
Chip Size	4mm x 4mm
Operating Frequency	MCVPU : 200MHz / others : 100MHz
Power Consumption	275 mW for full operation 216 mW for 3D Graphics 194 mW for image processing
Gate Counts	1.46M gates
Performance (Peak)	**Bandwidth** - 1.6Gbit/sec/pin [800MHz DDR2 DRAM] **MCVPU** - 38.4GOPS (16bit) [1.6G Pixels/sec] **UFU** - TU: 800M Texels/sec - FU: 400M Filtering/sec [any types] 400M Min_Max/sec **US** - 1.6GFLOPS [25M Polygons/sec]

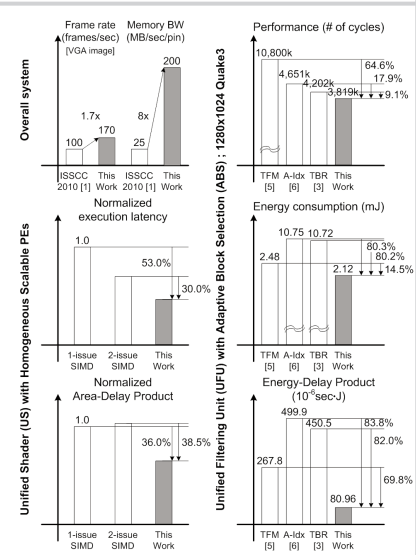

Figure 7.3.6: Chip specification and performance comparison.

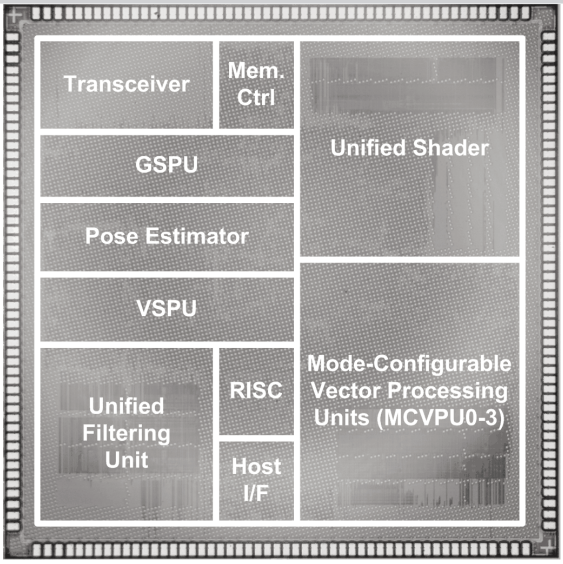

Figure 7.3.7: Chip micrograph.

978-1-61284-302-5/11/$26.00