A Scalable GPU Architecture based on Dynamically Reconfigurable Embedded Processor

*Won-Jong Lee, Sang-Oak Woo, Kwon-Taek Kwon, Sung-Jin Son, Kyoung-June Min, Gyeong-Ja Jang, Choong-Hun Lee, Seok-Yoon Jung, Chan-Min Park, Shi-Hwa Lee
Embedded Multimedia Systems Group, System Architecture Lab.
SAIT, Samsung Electronics

1. Introduction

The increasing cost of ASIC has been driving designers to choose more flexible solutions, as new chip architectures should be able to deliver high performance while maintaining low power consumption, area, and costs in shorter time-to-market environments. As a result, reconfigurable processors (RPs) have become increasingly important in recent years.

In this work, we present a new approach that utilizes an RP to design a tile based GPU, a core component of today's mobile application processors. Experimental results show that the use of RP based GPUs can be a versatile graphics solution, as they support scalable, flexible architecture models.

2. SRP based Scalable GPU Architecture

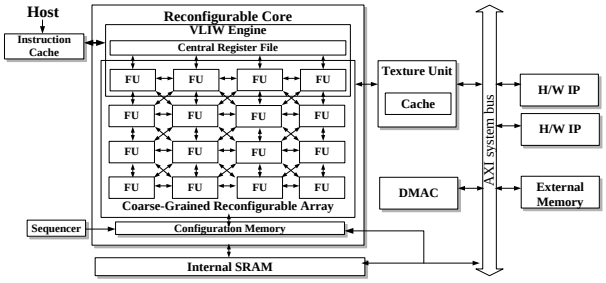

Fig 1. SRP system architecture

Figure 1 shows a SRP core which is a key component of our GPU architecture. The Samsung Reconfigurable Processor (SRP) is a proprietary low-power DSP core developed by Samsung Electronics, Ltd., as an elaboration on ADRES [Bingfeng et al 2005]. The SRP is a flexible architecture template that allows a designer to easily generate different instances by specifying different configurations in the target architecture.

The SRP includes a tightly coupled VLIW engine and a coarse grained reconfigurable array (CGRA). The VLIW engine is useful for general purpose computations such as function invocation and branch selection. The CGRA makes full use of the software pipeline technique via modulo scheduling to allow loop acceleration. Consequently, the CGRA exhibits a high IPC rate, up to the maximum number of FU arrays. During the execution phase, a central pointer allows dynamic reconfiguration of the FUs within a cycle. Instruction sets for executing shader kernels such as arithmetic, special function and texture operations are properly implemented in each FU. The integration of VLIW and CGRA allows the SRP to more effectively exploit instruction- and loop-level parallelism.

Figure 2 shows an instance of our multi-core SRP based GPU. Although the current architecture consists of one vertex and four pixel shaders, other configurations (e.g. multiple unified shaders) can be easily implemented using the SRP's reconfigurable feature. Tile based rendering (TBR) is adopted for power efficient architecture as the commodity embedded GPU. Each SRP core corresponds to a vertex and a pixel shader. Other units, requiring the least programmability as special purpose hardware for TBR, are embedded inside. Specifically, these units are batch management units (BMUs), primitive assembly and tile binning (PATB) units, tile dispatch units (TDUs), fragment generators (FGs) and raster operators (ROPs). Texture units are implemented as intrinsic in the SRP core and have two-level cache architecture. All components are connected to a 64-bit AXI system bus. The SRP and the hardware units are executed in a fully pipelined manner to exploit the high throughput rate of this arrangement.

*e-mail: joe.w.lee@samsung.com

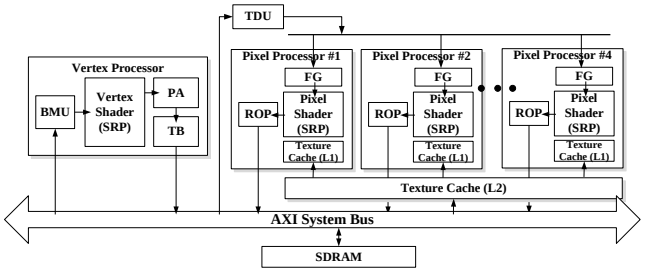

Fig 2. Our multi-core SRP based GPU architecture

TDU was designed for load-balanced parallel rendering. The TDU reads the triangles (tiles) from SDRAM in a swizzled order (e.g. space filling curve) and dispatches them to an idle pixel processor. This dynamic coherent scheme allows our GPU to achieve both load balancing and superior L2 cache utilization.

3. Experimental Results

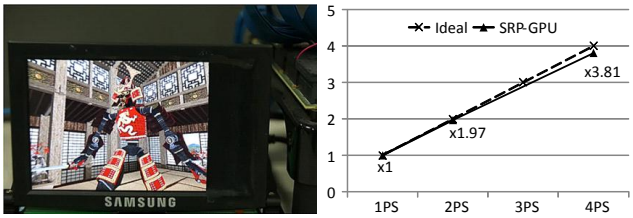

Fig 3. Rendering on FPGA (left) and performance result (right)

Our GPU is verified and evaluated by examining its performance during cycle-accurate simulation, verilog RTL simulation, and FPGA targeting. The SRP and the hardware units were synthesized to perform targeting on a Xilinx Virtex5 LX330 FPGA board running at 25 MHz with single port SDRAM modules and an LCD screen. Because of size limitations, the current FPGA implementation could equip itself with only two pixel shaders;, therefore, RTL simulation is used only for evaluating the performance for different numbers of pixel shaders. Cyber Samurai from Rightware's OpenGL|ES benchmarks has been thoroughly tested (Figure 3, left). Other various benchmark suites such as 3DMark Mobile ES 2.0, GLBenchmark are under testing.

Fig. 3 (right) shows a comparison of the performance result as a function of the scaling of the number of pixel shaders. The very effective use of pipelining and parallel rendering could allow our GPU to scale well, up to x3.81 with 4 pixel shaders. Our current GPU is predicted to render at up to 89.6 fps with WVGA (800x480) resolution at a 333 MHz core clock speed.

To achieve better performance with this device, we are currently redesigning the architecture to support more advanced features such as a unified shader or a multithreaded streaming CGRA. Finally, we expect that our GPU will be a core intellectual property for future application processors.

References

BINGFENG, M., ANDY, L., DIEDERIK, V., JEAN-YVES, M., AND RUDY, L., 2005. Architecture Exploration for a Reconfigurable Architecture Template. In *IEEE Design & Test of Computer*, vol. 22, No. 2. 90-101.