High performance memory accesses on FPGA-SoCs

a quantitative analysis

High Performance Memory Accesses on FPGA-SoCs: A quantitative analysis

Matthias Göbel, Chi Ching Chi, Mauricio Alvarez-Mesa and Ben Juurlink
Embedded Systems Architecture Group
Technische Universität Berlin
Berlin, Germany
Email: {m.goebel, chi.c.chi, mauricio.alvarezmesa, b.juurlink}@tu-berlin.de

FPGA-SoCs like Xilinx's Zynq-7000 and Altera's Generation 10 SoCs provide an integrated platform for HW/SW-codesign applications. Computationally complex tasks can be implemented in the programmable logic part while control logic is implemented on the CPU. A potential bottleneck in such approaches is the interface latency and the data transfer throughput. Especially the data transfer to and from the memory subsystems can decrease the achievable performance significantly. Therefore, an analysis of the according subsystems of the Zynq-7000 has been performed in order to estimate the possible performance of HW/SW-codesigns with a special focus on two-dimensional memory accesses.

Sadri et al. [1] have shown that the Zynq-7000 allows for a fullduplex throughput of up to 1708.5 MB/s between memory and programmable logic when running at 125 MHz. However, the work has some limitations like an exclusive focus on one-dimensional memory accesses and the use of only a single port (while the Zynq-7000 offers multiple of them). Furthermore, a relatively small and slow chip, the *XC7Z020*, has been used, resulting in a low bandwidth.

A data transfer engine that is able to read and write from/to main memory has been implemented. It can perform two-dimensional memory accesses with an arbitrary width, height and stride. The engine splits the overall transaction into smaller ones according to the AXI specification and generates the required control signals. By duplicating the engine, it is possible to use multiple memory ports of the Zynq-7000 in parallel thus increasing the overall bandwidth.

A synthetic benchmark for two-dimensional accesses has been designed in order to give an overview of the bandwidth that is achievable for given widths and heights. Some results for fullduplex access can be seen in Figure 1. By using all four *High Performance* ports of the Zynq-7000 in parallel, a throughput of 3440 MiB/s@214 MHz can be achieved when using the *XC7Z045* chip, with a theoretical maximum of the used DDR3-1066 chip of 4066 MiB/s. A cached, SW-based baseline reaches a maximum of 978 MiB/s@800MHz. Surprisingly, the coherent ACP achieves a relatively low maximum throughput of 159 MiB/s. Further analysis has shown that by reducing the height, a throughput comparable to the cached SW implementation can be achieved.

Furthermore, a benchmark has been designed that uses a trace of the memory accesses of the motion compensation stage of an HEVC decoder as an input and performs them. Again, the HW-based solution outperforms the SW-based one with a maximum throughput of 1515 MiB/s and 162 MiB/s, respectively, with the ACP solution showing a performance comparable to the SW implementation.

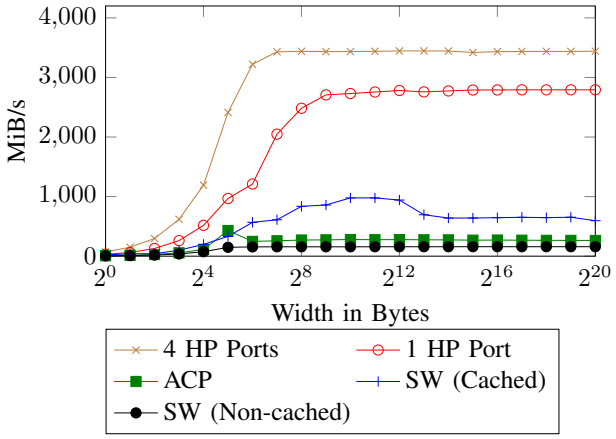

Figure 1: The fullduplex throughput for a fixed stride of 1 MiB and a height of 50 rows. The HP port solutions easily outperform the other solutions.

To conclude, the work has shown that, by carefully choosing a proper memory interface, a design in the programmable logic of a Zynq-7000 can untap a lot of the potential memory bandwidth. Therefore, it easily outperforms SW-based baselines and shows the potential of HW/SW-codesign approaches. Especially the combined use of multiple memory ports to increase the effective bandwidth proved to be a reasonable solution for expensive applications in terms of memory bandwidth like video decoders. In contrast, the coherent accesses offered by the ACP don't come for free and have therefore a limited scope of application. Future work in this area involves the use of alternative SoCs like Altera's Generation 10 SoCs and the application of real-world benchmarks from other domains than video coding.

References

[1] M. Sadri, C. Weis, N. Wehn and L. Benini, *Energy and Performance Exploration of Accelerator Coherency Port Using Xilinx ZYNQ*, ACM 10th FPGAWorld Conference, Copenhagen/Stockholm, Denmark/Sweden, 2013.