MP-STREAM: A Multi-Platform FPGA-Centric Memory Performance Benchmark

Syed Waqar Nabi
School of Computing Science
University of Glasgow
Glasgow G12 8QQ, UK
syed.nabi@glasgow.ac.uk

Wim Vanderbauwhede
School of Computing Science
University of Glasgow
Glasgow G12 8QQ, UK
wim.vanderbauwhede@glasgow.ac.uk

1. INTRODUCTION

We present MP-STREAM, an OpenCL-based multi-platform benchmark for sustained memory bandwidth. While our benchmark is heterogeneous, our focus is on FPGAs. The benchmark is based on the STREAM benchmark that has become the de-facto standard for CPUs[1], and has been ported to OpenCL for GPUs[2]. This benchmark has been developed in the context of our *TyTra* project on developing an optimizing compiler for running scientific code on FPGAs[3], which requires an estimate of achievable memory bandwidth.

2. BENCHMARK DESIGN

Our key contribution is the introduction of various generic as well as device-specific parameters that can be varied to measure their effect on sustained memory bandwidth. These parameters reflect both application and program characteristics. The following parameters can be varied in our tool which then emits custom OpenCL code and build scripts: target device (CPU, GPU, FPGA), choice of kernel (*copy, scale, add, triad*), data to/from device's DRAM (which is the main use-case) or directly from the host, type of word, size of array, degree of vectorization (i.e. memory-access-coalescing), data-access pattern (contiguous or strided), kernel-loop management (1 work-item or *NDRange* work-item kernel), flat or nested looping, loop unroll factor, work-group size, number of *SIMD* work-items, number of *compute-units*, pipelining options, and size of memory-ports.

One can see that these parameters constitute a significant FPGA-specific extension on the previous benchmarks designed for CPUs and GPUs.

3. RESULTS AND DISCUSSION

We experimented with four heterogeneous devices[1]. As an illustration, the result of one experiment where we vary the vectorization is shown in Figure 1. One apparent observation from our experiments was that OpenCL is not always performance portable across heterogeneous devices. Target-specific domain expertise or smarter high-level heterogeneous programming frameworks are thus needed for getting the best memory performance out of each architecture. For memory-bound applications – and high-performance computing (HPC) applications on FPGAs tend to be memory-bound – this memory performance becomes the overall performance determinant. We have made the case that HPC on FPGAs requires an extension in the available memory performance benchmarks, as there are a number of tuning parameters that effect FPGA memory bandwidth. Our contribution is a highly parametrizable benchmark specially tuned for FPGAs. The benchmark is publicly available[2].

[1]Intel Xeon E5, GeForce GTX Titan Black, Nallatech PCIe-385 with Altera Stratix V (aocl), Alpha-Data ADM-PCIE-V7 with Xilinx Virtex-7 (sdaccel).

Acknowledgments

The authors acknowledge the support of the EPSRC for the TyTra project (EP/L00058X/1).

4. REFERENCES

[1] John D. McCalpin. Memory bandwidth and machine balance in current high performance computers. *IEEE Computer Society TCCA Newsletter*, pages 19–25, December 1995.

[2] Tom Deakin and Simon McIntosh-Smith. GPU-stream: Benchmarking the achievable memory bandwidth of graphics processing units. In *IEEE/ACM SuperComputing,*, Austin, TX, USA, 2015.

[3] S. W. Nabi and W. Vanderbauwhede. Using type transformations to generate program variants for FPGA design space exploration. In *ReConfig 2015*, pages 1–6, Dec 2015.

[2]https://github.com/waqarnabi/mp-stream

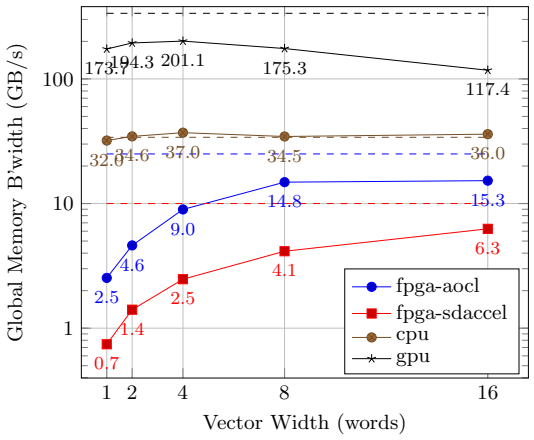

Figure 1: Varying vectorization for *copy* kernel on all targets. Array size is fixed at 4MB. Word size is 32 bits, and data is accessed contiguously in memory. No other optimizations are used.