Understanding How OpenCL Parameters Impact on Off-chip Memory Performance of FPGA Platforms

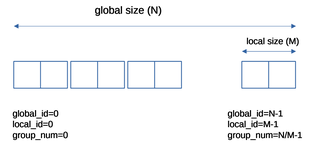
Yingyi Luo[1], Zheming Jin[2], Kazutomo Yoshii[2] , Seda Ogrenci-Memik[1]
[1] Northwestern University [2] Argonne National Laboratory

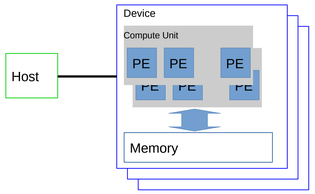
Abstract

Reconfigurability has strong potential to achieve higher performance and energy efficiency in the post-Moore era. Field-programmable gate arrays (FPGAs), the most practical reconfigurable architecture today, are becoming more relevant to scientific computing thanks to hardened floating-point circuits and emerging FPGA high-level synthesis (HLS) technology. Most notably, FPGA vendors started supporting OpenCL for FPGA platforms, and some OpenCL-based codes have been ported to FPGAs. However, OpenCL offers no guarantee for performance portability; optimal OpenCL parameters such as global size and local size are different between platforms, which could lead to unfair comparisons. In this study, our objective is two folds: 1) to understand how OpenCL parameters impact off-chip memory access performance of the current generation of OpenCL-FPGA platforms and 2) to find effective OpenCL parameters empirically from microbenchmark results.

OpenCL

- OpenCL is an open and royalty-free standard for general purpose parallel programming on heterogeneous computing
- Kernel: a function executed on OpenCL devices (e.g., FPGA, GPU)
- Compute unit (CU): comprises processing elements and memory interface
- Work-item: a set of parallel execution of a kernel
- Work-group: a group of work-items that execute on a compute unit
- Local work size: the number of work-items in a work-group

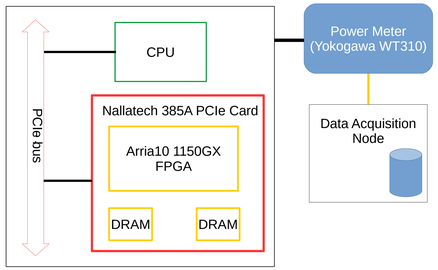


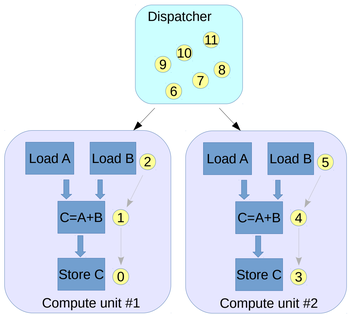


Experimental Setup

Target FPGA Platform:

- Nallatech 385A PCIe accelerator card
  - Arria 10 1150 GX FPGA
    - 34GB/s Memory bandwidth
    - Up to 1.5 TFLOPS
    - Idle power (~27 Watts)
    - 1,150K equivalent logic elements
  - 8GB DDR3 on-card memory
  - PCIe Gen3X8 Host Interface
  - OpenCL tool flow support



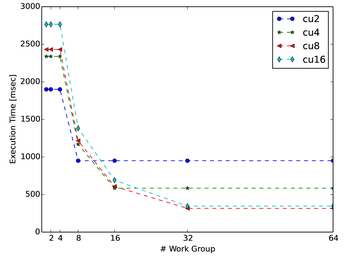


Streaming Memory Access

- A single-precision floating-point vector add microbenchmark
- A common regular memory access pattern in many HPC workloads
- The size of each vector is 512M (6GB of memory usage in total)

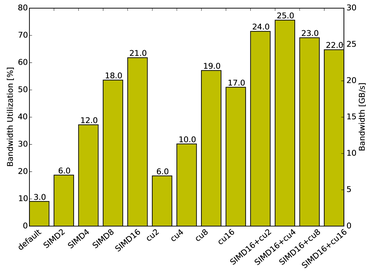
Impact of Work-group sizes

- The FPGA performance reaches the best when the number of work-group is above 32
- Kernel duplication may cause the global memory load/store operations from multiple compute units to compete for the global memory bandwidth



Impact of SIMD and Compute Units

- Kernel vectorization is better than kernel duplication for vector add
- Kernel duplication may cause the global memory load/store operations from multiple compute units to compete for the global memory bandwidth
- The combination of SIMD and compute units duplication can achieve the highest bandwidth

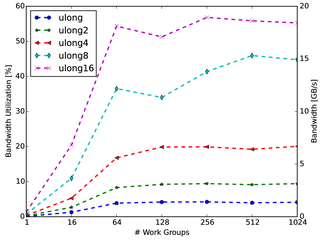


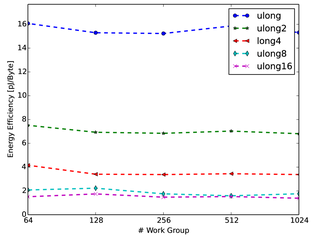
Irregular Memory Access

- A random memory access microbenchmark
- An irregular memory access found in pattern random number generation, binary search
- The data size is 6GB in total

Impact of Data types and Work-group sizes

- Small work-item size underutilizes the memory bandwidth
- 128 or more work-items maximize the bandwidth utilization
- Small data size cannot fully utilize the data bus
- Eight 64-bit (ulong8) accesses coalesced





Conclusion

- Designed microbenchmarks to understand how OpenCL parameters impact off-chip memory performance
- Measured the bandwidth and energy efficiency of two distinct memory access patterns on an Arria10-based board
- Finding the right global size and local size is the first important step; the single work-item kernel performs poorly on both VectorAdd and RandomIndex (up to ~60x slower than the best parameters)
- Larger SIMD size and larger data type, which yield wider memory operations, improve both memory bandwidth utilization and energy efficiency
- FPGAs design can achieve 75% (VectorAdd) and 55%(RandomIndex) of the memory bandwidth compared with the theoretical peak

U.S. DEPARTMENT OF ENERGY
Argonne NATIONAL LABORATORY
Northwestern University
Argonne National is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

SC17

Acknowledgments
This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-ACO2-06CH11357.