

Using Xilinx Alveo accelerators for Lattice QCD

Salvatore Calì, Grzegorz Korcyl, Piotr Korcyl



Asia-Pacific Symposium for Lattice Field Theory
August 4, 2020

Intro & Motivations

Towards Exa-scale Supercomputing

- mostly *traditional* architecture: CPU + # GPU
- no.1 FUGAKU supercomputer is based on ARM CPU nodes
- chinese Sunway is based on SW26010, a 260-core manycore processor
- no.1 HPCG TOP500 gives 3% of RMAX!!

Review & Status

Hardware installation attempts

- U. of Tsukuba: *Cygnus is the world first GPU-FPGA equipped supercomputer to be opened for public use ...*
- U. of Paderborn: *Noctua* supercomputer & multiple clusters

LQCD-related implementation attempts

- Trinity College Dublin, U. of Frankfurt & Maxeler, Jagiellonian University: conjugate gradient algorithm for Dirac operator

Hardware advantages

- complete system generation
- vast amount of logic resources
- High Bandwidth Memory
- UltraRAM blocks
- dynamic reconfiguration

Conceptual advantages

- data-oriented programming
- natural parallelism
- pipelining (computations/kernels)
- user-defined data-types
- kernel-to-kernel communication

Hardware setup: Xilinx Alveo accelerators: U250, U280, U50

- accessible in the Nimbix cloud
- access to the Zurich ETH Xilinx cluster
- Vitis 2020 environment
- openCL framework

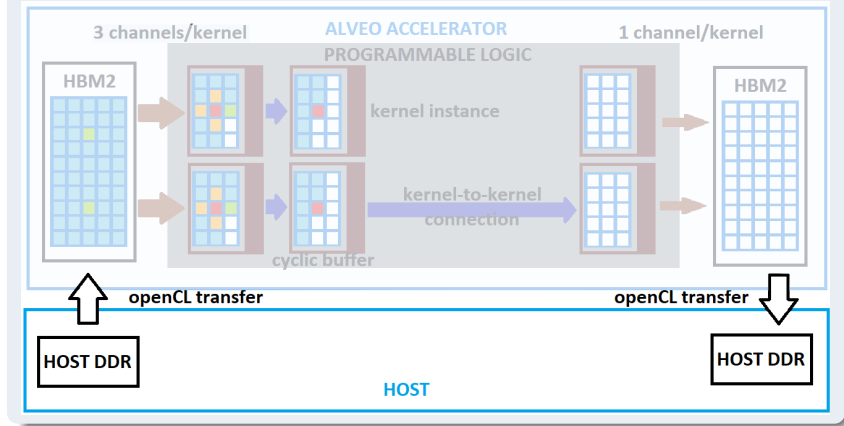
Relevant resources

- U50: 75 W, UltraScale+, 800k LUT, 8GB HBM2, 32 AXI channel access, PCIe Gen4 interconnect
- U250: 225 W, UltraScale, 1300k LUT, 64GB DDR, PCIe Gen3 interconnect
- U280: 225 W, Ultrascale+, 1300k LUT, 32GB DDR, 8 GB HBM2, PCIe Gen4 interconnect

Resource consumption

Single precision kernel: 756 BRAM, 1717 DSP, 149k LUT
latency 163 clock cycles

Implementation architecture



FPGA: Conjugate gradient algorithm

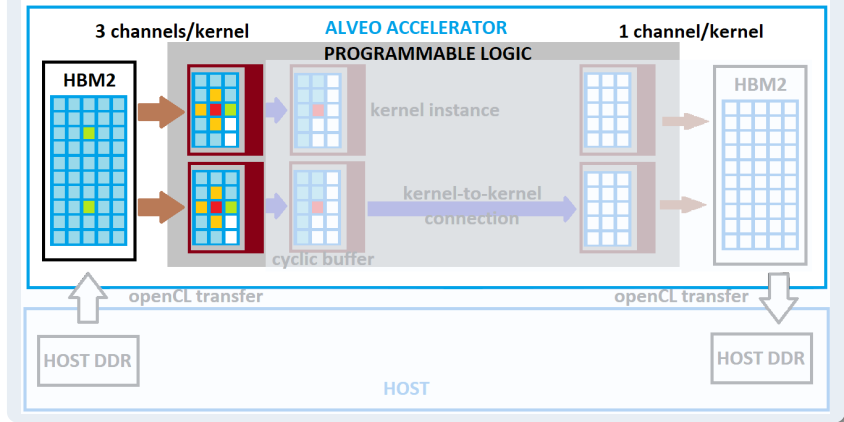
Main concepts

- one instance of kernel evaluates one stencil [KK18]
- data-streaming from HBM2: continuous usage of resources [KK19]
- pipelining computations: hiding latency, 2 clock cycle Initiation Interval [KK18]
- cyclic buffers: reduce data duplication, increases data reuse, stored in UltraRAM, offer kernel pipelining [SHY14]
- kernel-to-kernel communication: pipelining solver's iterations
- multi-node implementation: MPI communication via host-to-host

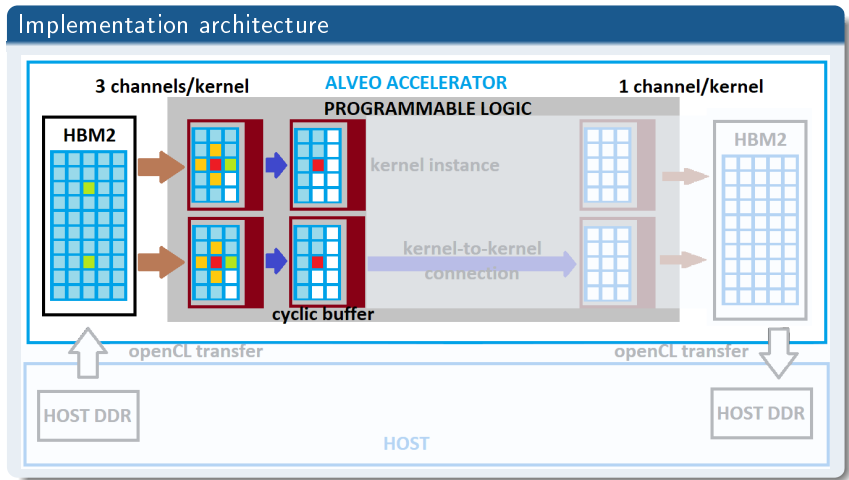
Details

- we use three 512 bit wide channels from HBM2 to transfer data (one spinor, 4 gauge links) in 2 clock cycles
- with data stored in buffers it is enough to evaluate the entire stencil
- local X,Y,Z dimensions limited by UBRAM size
- no restriction on the T dim. (local lattice has to fit into HBM2)

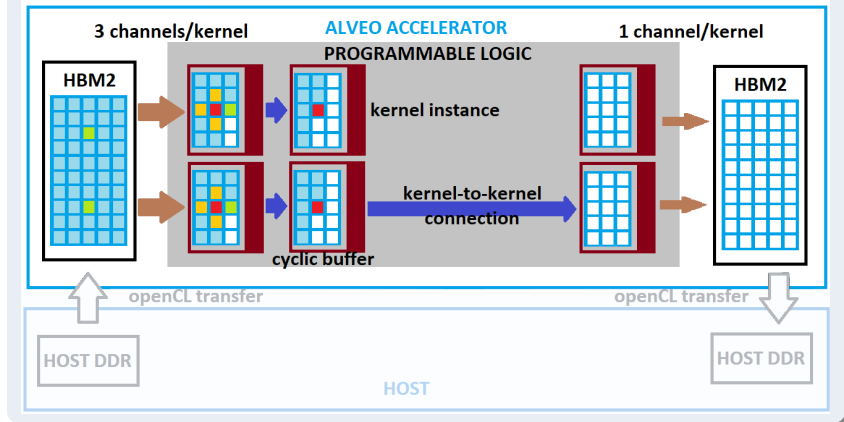
Implementation architecture



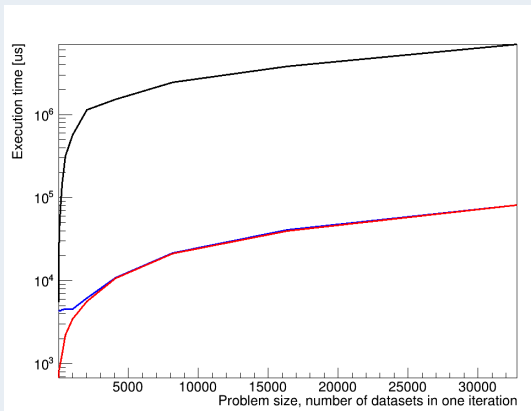
Implementation architecture



Implementation architecture



Runs details: Conjugate Gradient

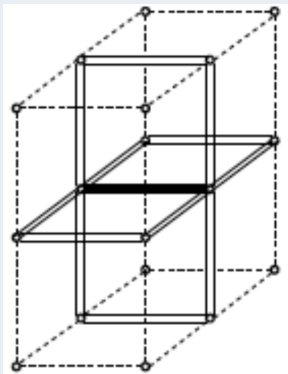


single precision performance: comparison between Alveo U280 and 1 core of Haswell CPU. Shown is the total number of lattice sites in the form of $4^3 \times T$. black - CPU-only, blue - FPGA+full memory transaction, red - FPGA-kernel only.

Dynamic reconfiguration

- multiple data types - dynamic reconfiguration between consecutive iterations allows to load a different kernel
- reconfiguration takes 70 ms
 - for the lattice with 32768 sites ($16^3 \times 8$):
 - one iteration in double precision takes 162 ms
 - one iteration in single precision takes 81 ms
 - for 2 or more iterations it's advantageous to reload
- gain even more significant for double-half data types

Another kernel: Smearing Procedure



Sketch of HYP smearing [HK01]

- other time consuming procedures can be delegated to accelerators
- smearing procedure with multiple iterations is the next candidate
- consecutive iterations can be pipelined in the kernel cascade
- timings similar to the CG: 2 clock cycles to read and store data in the cyclic buffers

Open-source code

Code available:

https://bitbucket.org/fpgafais/hpcg_fpga

Description:

<https://arxiv.org/pdf/2001.05218.pdf>

submitted to Elsevier SoftwareX





Prospects

- build local prototype with ≥ 4 Alveo cards
- benchmarking timings/power for different configurations
- HBM2 avoiding stream via PCIe from host memory
- further software development

Conclusions

- waiting for hardware to build a prototype
- many open research directions
- if you are interested, join us!

This work was in part supported by Deutsche Forschungsgemeinschaft under Grant No.SFB/TRR 55 and by the polish NCN grant No. UMO-2016/21/B/ ST2/01492, by the Foundation for Polish Science grant no.TEAM/2017-4/39 and by the Polish Ministry for Science and Higher Education grant no. 7150/E-338/M/2018. The project could be realized thanks to the support from Xilinx University Program and their donations.

-  [Anna Hasenfratz and Francesco Knechtli.](#)
Flavor symmetry and the static potential with hypercubic blocking.
Physical Review D, 64(3), Jul 2001.
-  [Grzegorz Korcyl and Piotr Korcyl.](#)
Towards Lattice Quantum Chromodynamics on FPGA devices.
Computer Physics Communications, 2018.
-  [Grzegorz Korcyl and Piotr Korcyl.](#)
Investigating the Dirac operator evaluation with FPGAs.
Supercomputing Frontiers and Innovations, 2019.
-  [Kentaro Sano, Yoshiaki Hatsuda, and Satoru Yamamoto.](#)
Multi-fpga accelerator for scalable stencil computation with constant memory bandwidth.
Parallel and Distributed Systems, IEEE Transactions on, 25:695–705, 03 2014.