Graph Analytics on RISC-V GPU: Where are the Bottlenecks?

Nimish Shah and Marian Verhelst, ESAT-MICAS, KU Leuven

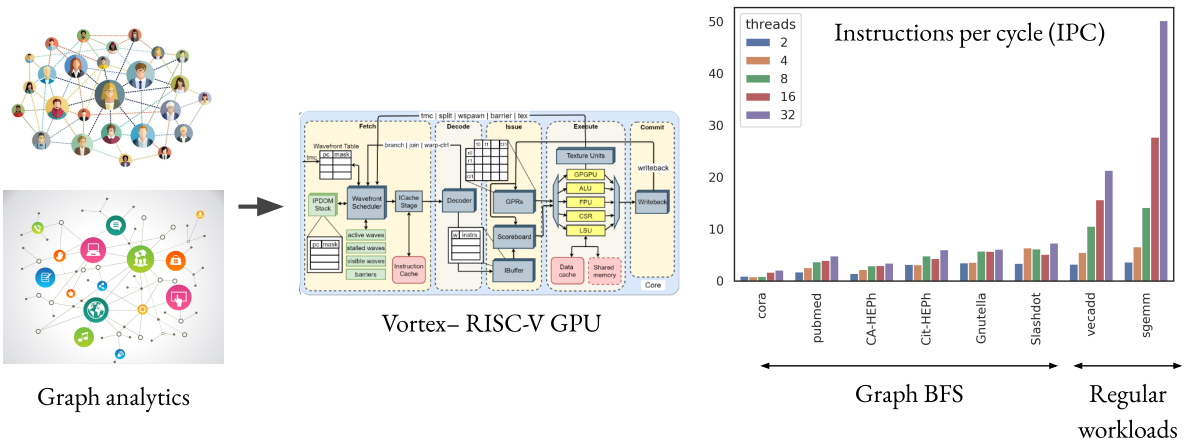
Abstract

RISC-V ISA extensions are actively researched by the open-source community for a wide spectrum of applications like machine learning, digital signal processing, and high-performance scientific workloads, for both energy-constrained IOT and performance-critical general-purpose platforms. Recently, to further extend the utility of RISC-V ISA to more applications, Tine et al. proposed Vortex – a RISC-V-based general-purpose graphics processing unit (GPU), developed by adding only 6 new instructions to the ISA. The aim of this GPU extension is to provide an open-source platform for fostering research on single instruction multiple threads (SIMT) based architectures, which target highly-parallel workloads that are not typically suitable for CPUs (like 3D graphics rendering). In this work, we benchmark the performance of a highly-parallel but *irregular* set of workloads from **graph analytics**, which are not usually suitable for GPUs, and compare it to regular workloads to identify the key bottlenecks in the GPU architecture.

Vortex – a RISC-V GPU. The Vortex GPU developed by Tine et al. [1] extends the RISC-V ISA with simply 6 new instructions to enable a datapath with a SIMT-based execution model. SIMT is suitable for highly parallel workloads that execute the same program on a large number of data elements as the datapath can use multiple (and simple) processing elements in parallel. The ISA also supports thread divergence which is a crucial feature that differentiates SIMT from SIMD. The open-sourced Vortex framework contains a compiler for OpenCL code, a cycle-accurate simulator, RTL verilog description, and an FPGA implementation. The architectural parameters like the number of parallel hardware threads, warps, cores, etc. can be easily configured.

Graph analytics. Graph, in which nodes are interconnected with edges, is a fundamental datastructure used for modelling road networks, social networks, chemical molecules, communication networks, etc. To extract meaningful insights from graphs, a variety of analytics algorithms are commonly used like breadth-first search, single-source shortest path, clustering. The graphs are usually large (more than 10k nodes/edges) and the algorithms exhibit significant parallelism to allow parallel hardware execution. Furthermore, in an *iteration*, the same (usually short) program is executed for all the *active* nodes or edges, making these algorithms suitable for GPU-like SIMT execution.

This work. Despite the fit with the SIMT execution model, graph analytics typically under-perform on GPUs. The chart (bottom right) demonstrates this for the breadth-first search (BFS) algorithm on Vortex. Despite the large size of the graphs and ample parallelism, the throughput does not scale like the regular workloads. In this work, we profile the performance of the different microarchitectural blocks of the GPU and identify the key hardware bottlenecks and the graph properties that cause these bottlenecks. Finally, we will propose mitigation techniques to alleviate these hardware bottlenecks. Thus, this work contributes to the ecosystem of an extendable RISC-V-based GPU that can be tuned depending on the workload characteristics.



[1] Tine et al., "Vortex: Extending the RISC-V ISA for GPGPU and 3D-Graphics", MICRO, 2021.