FPGA as a RESTful Service: Release Your Accelerator From the PCIe Cage!

Fabio Maschi  
Systems Group, Dept. of Computer Science  
ETH Zurich, Switzerland

Gustavo Alonso  
Systems Group, Dept. of Computer Science  
ETH Zurich, Switzerland

ABSTRACT
FPGAs have proven to be an efficient execution node to accelerate compute-intensive tasks in distributed system. However, keeping them as a primary PCIe-attached device remains a limiting factor for their broader adoption: (i) the tight coupling of technology stacks increases development and maintenance costs; (ii) PCIe communication pattern reduces the potential for low-latency and high-bandwidth acceleration; and (iii) cloud offerings are restricted to few settings. In this work, we present Strega, an HTTP server-side stack for FPGA kernels, making them available as a RESTful micro-service directly over the network. The enhanced hardware abstraction facilitates software integration, while offering deterministic, orders of magnitude higher performance.

1 INTRODUCTION
The PCIe bus has been the traditional interconnect for accelerators, such as FPGAs, GPUs and TPUs. It exposes the hardware lanes to the host CPU and operates with PCIe transactions (i.e., data read/write, interruptions, configuration). To issue a full kernel invocation, the host needs to (i) allocate memory in the device; (ii) transfer the input data to the device’s memory; and (iii) trigger the start of the kernel. The accelerator then (iv) fetches the input data; (v) executes it; (vi) writes the result data back into its memory; and (vii) no-