Hyper-acceleration is a technology that enables big data and machine learning applications to automatically utilize the power of unconventional hardware - GPUs, FPGAs - as well as software optimizations with many-core CPUs.

The Bigstream Hyper-acceleration Layer functions as a runtime system that sits between a software platform (such as Apache Spark, or TensorFlow) and the underlying hardware to slice and distribute the computation between traditional CPU cores and different accelerator resources like FPGAs and GPUs.

These results were generated with 79% logic utilization and 15% Block RAM utilization on an Arria 10 FPGA compared to results from an Intel Core i7-5930K server (6 cores, 12 threads, 12 logical cores) running Spark 2.1.1 with 10 Spark executors.

We achieve these results by mapping the set of operators in a stage of a Spark query plan to physical hardware microarchitectures, in the form of bit files for the FPGA. We use a stage acceleration analyzer in our Gorilla++ streaming compiler to analyze the physical plan of a query and to decide which slice of the dataflow is optimally executed on FPGA, CPU, or GPU. At run-time, the optimal bit file for this dataflow stage is chosen from a cache of bit file templates.