Tradeoffs and Considerations in the Design of Accelerators for Database Applications

Roger Moussalli
Accelerator Platforms Group
IBM TJ Watson Research Center
Yorktown Heights, USA
rmoussal@us.ibm.com

Abstract— **General purpose processors have traditionally been favored over application-specific architectures due to the provided flexibility, standardized and simpler programming model, as well as significant reduction in development time. Fueled by the steady advances in transistor scaling, general purpose CPUs satisfied the performance needs of most applications.**

While CPUs were becoming ubiquitous, advances in digital storage technologies and sensing devices (cameras, microphones, etc) led to massive and sky-rocketing amounts of data being generated by devices of all scales. Extracting insights out of this Big Data introduces significant opportunities for business intelligence, though the growth of data volumes and complexity of query patterns has been increasing at a startling rate. With Moore's law ending, transistors' shrinking coming to a halt and CPU performance saturating, accelerator technologies are increasingly embraced to augment general purpose CPUs and to address performance concerns.

Accelerators diverge from traditional CPU architectures in the way they utilize the available silicon resources. In particular, accelerators maximize the resources available for raw computing (ALUs, Floating Point Units) and push back the burden of correct program semantics and control to higher levels of the stack including the compiler and programming models, while focusing on a selected subset of applications.

Accelerators include Application Specific Integrated Circuits (ASICs), Field Programmable Gate Arrays (FPGAs) and General Purpose Graphics Processing Units (GPGPUs), each with their programming model, advantages and challenges. FPGAs enable the deployment of deep custom pipelines, whereas GPGPUs provide hundreds of small processors executing in a massively parallel fashion. Compared to CPUs, accelerators attain higher performance out of the available transistors for a wide range of applications.

This talk covers tradeoffs of accelerators (FPGA and GPGPU) specific to a set of database applications, namely XML filtering, spatiotemporal analytics in the context of the Internet of Things, and relational database querying. Tradeoff metrics include programmability, performance, accuracy and energy consumption.

While accelerators achieve high speedups for "hot" code paths, the attach point of accelerators in a system significantly impacts the end-to-end application performance. As such, system and deployment-level considerations must be made. To this end, I will go over IBM's efforts to facilitate the inclusion and increase the adoption of accelerators. These include (1) the Coherent Accelerator Processor Interface (CAPI), reducing software refactoring from the CPU side as well as CPU-accelerator latency; (2) the ConTutto research platform for acceleration innovation in the memory subsystem, providing very high bandwidth to accelerators; and (3) NVLink®-enabled IBM POWER® processors.

Keywords—accelerator; acceleration; hardware; FPGA; GPU; database; ConTutto; CAPI; NVLink