Power has become one of the key challenges impacting the design of new generations of high performance computing systems. Technological trends force a growing part of the power budget to be spent on data transfer and the memory system, while only a fraction is used for actual computation. The impact of these developments is further stressed by new workloads such as Big data analytics, which process huge amounts of data that often exhibit limited locality and, consequently, increase the amount of data that has to be transferred throughout the memory hierarchy.

In order to cope with these problems, computer systems are increasingly being designed with particular workloads in mind. Such workload-optimized systems typically try to optimize the hardware and software stack in a holistic fashion to achieve the best performance per Watt for the target applications. Another interesting development is the move from the conventional compute-centric model to a more data-centric one in order to reduce expensive data movements by "bringing the computation closer to the data."

Triggered by the above trends, the Accelerator technologies group at the IBM Zurich research laboratory is investigating hybrid computer systems that offload part of the processing to programmable near-memory accelerators that tightly integrate computation with main memory operation. Besides exploiting the data-centric concept to limit data transfers, this research targets to further improve power efficiency by letting the memory system play a more important role in (dynamic) workload optimization by enabling different ways to adapt the memory system operation to the workload behavior, and by exploiting a novel architecture and programming model to reduce programmability and related overhead compared to traditional architectures.

For demonstrating and investigating the capabilities of our concept, we are currently building a demonstrator based on OpenPOWER hardware, which includes POWER8 processors, FPGAs and GPUs, together with the corresponding tool environment.

We are looking for students that are interested to do a graduation project as part of this project. There is a range of topics available related to compilation and runtime optimization, the implementation and performance evaluation of specific applications (e.g., stencil processing, analytics, sparse matrix, compression), as well as architectural and implementation related aspects (e.g., exploitation of dynamic partial FPGA configuration). Please contact us if you are interested.