EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

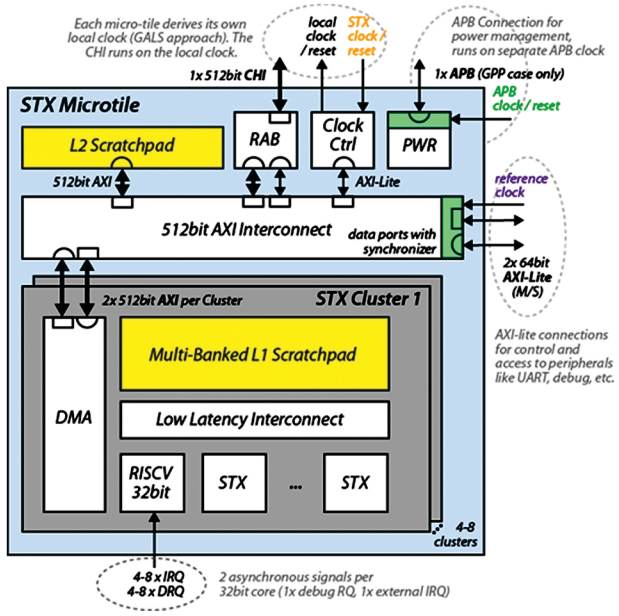
Domain specific accelerators in EPI
STX (stencil/tensor accelerator)

Architecture

- Domain specific accelerator for Machine Learning and Stencil operations within EPAC tile
- Designed to provide at least 10x more energy efficient computing for these workloads
- A typical STX accelerator has 4-8 clusters where each has
  - multiple specialized compute units
  - one 32bit RISC-V processor for support
  - energy efficient local scratchpad memory
  - DMA to transfer data from/to the accelerator

Goal: at least 5x more energy efficiency (TFLOPS/W)

Description



From the beginning EPI explicitly considered “specialised blocks for stencil and deep learning (DL) acceleration. The vector and stencil capabilities will address workloads in HPC centres, while the DL block will target learning acceleration” as part of the acceleration stream motivated by “optimised performance and energy efficiency” for “specialised computations”. In the initial DoA, two different domain specific accelerators (NTX for machine learning, and a stencil accelerator) were suggested. During the first few months of the project, researchers from Fraunhofer Institute, ETH Zürich and University of Bologna were able to merge the functionality of both units into a very efficient computation engine that has been named STX (stencil/tensor accelerator).

Such ”domain-specific accelerators” are now a major trend in industry, as can be seen by multiple new announcements in the 2019 Hot-Chips symposium and AI Summit by industry heavyweights as a multitude of startups that have presented acceleration engines that were based on specialized datapaths and not general purpose vector units, confirming the significant differentiation in architectures needed for achieving top efficiency and performance in the machine learning domain.

The main goal of STX is to achieve a significantly higher (at least 5x-10x) energy efficiency over general purpose/vector units. The efficiency tells us how many computations can be performed with the unit, and the early target for the STX unit was to achieve at least 5x more energy efficiency (TFLOPS/W) than the vector unit on deep learning applications. In the first few months of the project, it became clear that these estimations are rather conservative, and the effective efficiency within EPI chips will be significantly higher. For applications that require only inference using quantized networks, this efficiency will be another 10x higher.

STX has been designed as a modular building block with several parametrization options. Each STX accelerator consists of several clusters of computing units, a typical instance would have four such clusters. Each cluster in turn consists of specialized computing engines as well as up to two RISC-V cores that are used to control the computing engines and perform additional operations. All these units will access a local scratchpad memory, which will be filled using a centralized DMA unit. This configuration allows for 64 GFLOPS (single precision FP), and multiple instances of STX can be instantiated in an EPAC tile.

STX is programmed using OpenMP, there are solutions that allow regular operations to be offloaded to the STX unit from an ARM system (in the GPP) or the 64-bit RISC-V core (in the EPAC tile) using both GCC and LLVM based flows that will be further refined as part of the project.

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826647