Contribution ID: **23** Type: **Standard Talk**

Running Converged HPC & AI Workloads on the Groq AI Inference Accelerator

Wednesday 27 September 2023 16:00 (15 minutes)

Converged compute infrastructure refers to a trend where HPC clusters are set up for both AI and traditional HPC workloads, allowing these workloads to run on the same infrastructure, potentially reducing underutilization. Here, we explore opportunities for converged compute with GroqChip, an AI accelerator optimized for running large-scale inference workloads with high throughput and ultra-low latency. GroqChip features a Tensor Streaming architecture optimized for matrix-oriented operations commonly found in AI, but GroqChip can also efficiently compute other applications such as linear algebra-based HPC workloads.

We consider two opportunities for using the Groq AI accelerator for converged HPC. The first example is a structured grid solver for Computational Fluid Dynamics (CFD). This solver can run in a classical implementation as a direct numerical solver (DNS) using the pressure projection method. In a hybrid AI implementation, the same DNS solver is augmented with CNN-based downscaling and upscaling steps. This enables a reduction of grid size from 2048 to 64, thus significantly reducing the amount of compute necessary while maintaining a similar quality of results after upscaling. A speedup of three orders of magnitude is made possible by the combination of reducing the number of compute steps in the algorithm through introducing AI, and by accelerating both the CNN and DNS stages with GroqChip. The second example is using HydraGNN for materials science and computational chemistry. These problems are typically solved with Density Field Theory algorithms, but recently, Graph Neural Networks (GNNs) have been explored as an alternative. For example, GNNs can be used to predict the total energy, charge density, and magnetic moment for various atom configurations, identifying molecules with desired reactivity. The computation requires many parallel walks of HydraGNN with low batch sizes, and can be solved on GroqChip 30-50x faster than an A100 graphics processor.

Authors: Mr ZHANG, Chenyu (Maxeler Technologies); Mr SHANMUGAVELU, Sanjif (Maxeler Technologies); Dr BECKER, Tobias (Maxeler Technologies)

Presenter: Dr BECKER, Tobias (Maxeler Technologies)

Session Classification: Contributed Talks

Track Classification: Contributed Talks