Optimising Network-on-Chip Architecture for Deep Learning Accelerators at Scale

Specialised deep learning accelerators such as AMD's VERSAL[1] family of devices are gaining traction in data centers and high-performance compute clusters. VERSAL architecture integrates dedicated AI engines which are hardware blocks specially designed to execute AI tasks at high throughput while consuming much lower energy than GPU-based methods. VERSAL platforms for HPC typically include multiple such AI engines which are interconnected by a network on chip (NoC) architecture, also connecting them to other resources on the same chip, such as the ARM cores, multi-Gigabit network interfaces (>100 Gigabit), DSP engines, and a general programmable logic area for custom digital designs[2]. This highly parallel architecture combined with flexible connectivity allows for these devices to deploy large-scale AI models, commonly used in applications such as post-production workflows, among others.

The key challenge in mapping large-scale AI models to such architectures, however, is to determine the best suitable interconnect configuration that minimizes data latency, congestion, and stalling[3,4,5]. The fully customisable NoC architecture on VERSAL combined with the ability to interface with other resources on the chip, such as the Network Interface or bespoke logic for data preprocessing on the programmable logic, creates a large design space to be explored to arrive at an optimal NoC configuration. Additionally, while the tools from AMD perform mapping of tasks to the accelerators, optimising data movement for maximal efficiency is left as a user-driven task, often leading to reduction in bandwidth between different parts of the system compared to the theoretical limits of the interface. This impacts not only the performance throughput achievable on such designs, but also the energy consumed per inference. As the complexity of AI models increases, improving the energy efficiency of AI inference is a key consideration. It is estimated that large language models services like ChatGPT consume about 1 GWh a day with hundreds of millions of requests[6] while complex AI models are evolving in the media industry to perform denoising at high resolution (4K and beyond at HDR), rendering visualisations and effects, segmentation and green screen keying, among others. Developing a design flow that can map the AI model efficiently on VERSAL (and similar heterogenous) platforms, that extracts maximal efficiency from the NoC, will hence have a significant impact on the performance of the AI model (in terms of frames per second for instance) and energy efficiency (in terms of mJ per inference).

This project will aim to develop a set of tools that can:

1. Benchmark and determine the limits of the NoC interconnect for a system architecture involving multiple blocks (such as the 100 Gigabit network interface, preprocessing blocks in logic)

---

[1] Versal ACAP System Integration and Validation Methodology Guide (UG1388), AMD

[2] Versal ACAP Adaptive SoCs Design Guide (UG1273), AMD

[3] Brown, Nick. "Exploring the Versal AI engines for accelerating stencil-based atmospheric advection simulation." Proceedings of the 2023 ACM/SIGDA International Symposium on Field Programmable Gate Arrays. 2023.

[4] Lang, Ian, Nachiket Kapre, and Rodolfo Pellizzoni. "Worst-case latency analysis for the Versal NoC network packet switch." IEEE/ACM International Symposium on Networks-on-Chip. 2021.

[5] Zhang, Chengming, et al. "H-gcn: A graph convolutional network accelerator on Versal ACAP architecture." International Conference on Field-Programmable Logic and Applications (FPL). IEEE, 2022.

[6] https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/

2. Arrive at the optimal parameters for the NoC configuration to meet the performance requirements
3. Perform energy-performance trade-off analysis for the specific application.
4. Investigate low-level optimisations such as representation formats and precision for data transferred through the NoC without sacrificing the accuracy of the AI model.

To achieve these goals, the project will leverage close collaborations with AMD's Versal architecture research group based in Dublin. For developing, testing and validating the tools developed, we propose to utilise AMD's HPC research platform, HACC, (hosted by ETH Zurich in Europe) which offers remote access to VERSAL devices as well as other high-end FPGA nodes for performing a comparative study. My research team at Trinity already has access to HACC and will facilitate access to the student undertaking this project.

Additionally, I am Trinity's Co-PI on the Horizon Europe Project, EMERALD, which investigates energy efficiency of AI-driven applications in post-production media pipelines. The research outcomes, data, and tools from EMERALD can be used to drive the AI use case in this project, when performing application-driven optimisation of the NoC and for energy-performance trade-offs.