UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

ENGINEERING

Initial Statement of Research Intent

Notes to the applicant:

- If your application progresses, you may be asked to submit a more detailed Statement of Research Intent.
- Applicants should note that students are expected to submit their thesis after 3-4 years full-time study (6-8 years part-time).
- Your project outline should be a minimum of 500 words or one page.
- Along with the other documents requested under the current system (including CV, transcripts, names of referees, demographic information, other sources of funding, confirmation goals).
- If you're interested in funding, please visit University of Auckland Doctoral Scholarships.

Your details

Name of applicant:	Lingyu Gong
Student ID (if known):	249655880
Department/s (if known):	Electrical and Electronic Engineering
Proposed supervisor/s (if known):*	

* Applicants can find academics who are accredited to undertake doctoral supervision on the University Website and are expected to contact potential supervisors. For more information, please see Discovery profiles.

Those unable to do so will be supported, but the application process may be delayed. If no appropriate supervisor can be identified, the application will be declined.

Area of research interest

Provisional title of thesis or area of research interest:

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

Background, including literature review:

Specialised deep learning accelerators such as AMD's VERSAL 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. This highly parallel architecture combined with flexible connectivity allows for these devices

Initial Statement of Research Intent

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. 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).

Significance and contribution to existing literature:

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).
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.

Methodology:

To achieve these goals, this project will work closely with AMD's Versal architecture research team. To develop, test and validate the developed tools, I plan to use AMD's HPC research platform HACC (hosted by ETH Zurich in Europe), which provides remote access to the VERSAL device as well as other high-end FPGA nodes for comparative studies.

Initial Statement of Research Intent

Expected outcomes (optional):

References/bibliography (not included in the word count):

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/

Research/professional background

Briefly summarise any previous research/professional experience in your area of interest

Undergraduate Research:

During my undergraduate studies at Capital Normal University, I focused on Network-on-Chip (NoC) systems. My bachelor's thesis, titled "Research on Open Source Network on Chip (NoC) based on Intelligent Routing Algorithm," involved exploring NoC concepts, topologies, and the XY routing algorithm. I used OPNET software to simulate and test this algorithm, analyzing its latency and throughput. Additionally, I investigated three open-source NoC generators: OpenSoC Fabric, OpenSMART, and Constellation, and conducted detailed testing of the Constellation generator within the Chipyard environment.

Master's Research:

In my master's program at Trinity College Dublin, my research aimed to optimize NoC performance prediction using AI techniques. Titled "Enhancing On-Chip Network Predictions with Advanced AI Techniques," the project involved configuring NoC networks using Booksim2, generating datasets, and developing AI models such as DNNs and CNNs to predict performance parameters like throughput and latency. This approach aimed to reduce simulation reliance and accelerate the design flow. The project successfully demonstrated improved prediction accuracy and efficiency.