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

**1. Objectives and Significance**

The project aims to develop a set of tools for optimizing the NoC architecture in deep learning accelerators, specifically AMD’s VERSAL devices. These tools will benchmark the NoC interconnect, determine optimal configuration parameters, analyze energy-performance trade-offs, and explore low-level optimizations to improve the overall performance and energy efficiency of AI models.

**2. Current State of Practice**

Currently, tools from AMD perform basic mapping of tasks to accelerators but leave the optimization of data movement to users. This often results in suboptimal bandwidth utilization and higher energy consumption. The lack of comprehensive tools for NoC optimization creates a large, unexplored design space, making it challenging to achieve the best performance and energy efficiency.

**3. Innovation and Success**

This research introduces a holistic approach to NoC optimization by developing tools that automate the benchmarking, configuration, and trade-off analysis processes. Leveraging close collaboration with AMD’s Versal architecture research group and utilizing the HACC platform for testing and validation, this project aims to provide precise and actionable insights for optimizing NoC configurations. The integration with the Horizon Europe Project, EMERALD, will further enhance the relevance and application of the research outcomes.

**4. Relevance and Impact**

This research is significant to data center operators, high-performance computing (HPC) clusters, and industries that rely on large-scale AI models for applications such as media post-production, natural language processing, and advanced simulations. Improved NoC configurations can lead to significant cost savings, better performance, and reduced energy consumption.

Successful optimization of the NoC architecture will result in higher performance (e.g., increased frames per second for AI models) and improved energy efficiency (e.g., lower mJ per inference). This will enable more effective deployment of large-scale AI models, enhancing their practical applications and reducing operational costs in data centers.

**5. Risks and Payoffs**

The main risks involve the complexity of accurately modeling and simulating the NoC and AI engine interactions. However, the potential payoffs include significant improvements in AI model performance and energy efficiency, which are critical for scaling AI applications in a cost-effective manner.

**6. Evaluation**

Midterm evaluations will include the successful development and initial testing of the benchmarking and configuration tools. Final evaluations will assess the performance improvements and energy efficiency gains in AI models deployed on VERSAL platforms, validated through comprehensive testing on the HACC platform.

**7. Appendix: Prior Research Experience**

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.

**8. Conclusion**

This research proposal aims to significantly enhance the performance and energy efficiency of deep learning accelerators by optimizing the NoC architecture. With the support of Johns Hopkins University's esteemed faculty and resources, I am committed to achieving these goals and contributing valuable insights to the field of high-performance computing and AI.