**Detailed proposal**

**\* Please provide details of your proposed research to include (a) aims, objectives and central research questions of the project, (b) how existing literature on the topic has been used to inform the proposal and (c) how the project will advance state of the art and make a contribution to existing knowledge: 500 words**

(a) Aims, Objectives, and Central Research Questions:

The primary aim of this research is to optimize the Network-on-Chip (NoC) architecture within AI accelerators, specifically AMD's VERSAL platform, to enhance performance and energy efficiency. The objectives include developing a set of tools and methodologies that can:

1. Benchmark the NoC interconnect to identify performance bottlenecks.
2. Determine the optimal NoC configuration parameters to minimize data latency and maximize throughput.
3. Perform energy-performance trade-off analysis for specific AI applications to balance speed and power consumption.
4. Investigate low-level optimizations in data transfer formats and precision without sacrificing the accuracy of AI models.

The central research questions guiding this project are:

How can the NoC configuration within VERSAL AI accelerators be optimized to reduce data latency and congestion?

What are the trade-offs between performance and energy consumption in different NoC configurations?

How can data representation formats be adjusted to enhance energy efficiency without affecting AI model accuracy?

(b) Existing Literature and Its Influence on the Proposal:

Current literature shows that NoC optimization plays a critical role in improving the performance of AI accelerators. Studies, including those on VERSAL AI engines and network packet switching, have identified the challenges in managing data movement within these architectures. However, there is a gap in user-driven optimization tools that can tailor the NoC for specific AI workloads. Research on energy efficiency in large-scale AI models, such as high-resolution media processing, emphasizes the need for hardware-level optimizations to achieve sustainable performance. This proposal builds upon these findings by aiming to develop practical tools that address the complexity of NoC configurations, directly influencing the efficiency of AI tasks at scale.

(c) Advancing the State of the Art and Contribution to Knowledge:

This project will advance the state of the art by providing a comprehensive framework for optimizing NoC architecture in AI accelerators. Unlike existing tools that focus primarily on task mapping, this research will delve into fine-tuning the NoC to achieve optimal performance and energy efficiency. By offering a detailed analysis of NoC configurations, energy-performance trade-offs, and data transfer optimizations, the project will contribute new methodologies for managing complex AI workloads. The collaboration with AMD's Versal research group and access to their advanced hardware will ensure that the proposed tools are validated in real-world settings. Ultimately, this research will make a significant contribution by offering a scalable solution for enhancing AI hardware performance, benefiting applications that demand real-time processing and lower energy consumption.

**\* Please detail the research design and methodologies to be employed in carrying out your scholarship which should be described in sufficient detail to demonstrate your thorough understanding of the research topic: 500 words**

This research aims to optimize the Network-on-Chip (NoC) architecture in AMD's VERSAL AI accelerators using a multi-phase methodology involving benchmarking, simulation, and optimization.

1. Benchmarking and Analysis:

The initial phase involves benchmarking the VERSAL NoC using AMD's development kits and custom profiling scripts. Key metrics such as data latency, congestion points, and bandwidth limitations will be identified under various AI workloads. This analysis will provide a baseline for understanding the NoC's performance and areas for improvement.

2. NoC Configuration Optimization:

Using simulation tools like BookSim and AMD's NoC modeling software, we will explore the design space of NoC configurations. Parameters such as routing algorithms, buffer sizes, and link widths will be adjusted to identify the optimal configuration for minimizing latency and maximizing throughput. Automated search techniques like genetic algorithms will aid in discovering the most efficient NoC setups.

3. Energy-Performance Trade-off Analysis:

This phase conducts a detailed analysis of energy and performance trade-offs using power modeling tools and AMD's VERSAL toolkit. Various NoC configurations will be assessed for energy consumption in relation to their performance. Techniques like dynamic voltage and frequency scaling (DVFS) will be explored to optimize energy efficiency during runtime.

4. Low-level Data Transfer Optimization:

To enhance efficiency further, we will investigate data encoding and compression techniques to reduce the size of data transferred across the NoC. Custom data representation formats will be tested to minimize bandwidth usage while preserving AI model accuracy. Experiments will include analyzing the impact of different data precisions on throughput and energy efficiency.

5. Validation and Real-world Testing:

The final phase involves validating the optimized NoC configurations on AMD's VERSAL hardware using the HPC research platform, HACC. Real-world testing with complex AI applications will measure performance metrics such as frames per second and energy per inference to validate the effectiveness of the optimizations.

**\* Please provide a schedule to include (a) milestones and deliverables for completion of the proposed research, (b) risks that might endanger reaching these deliverables and (c) the contingency plans to be put in place in order to mitigate these risks: 500 words**

In Year 1, the research will focus on establishing a strong foundation. The first quarter will involve a comprehensive literature review of NoC architecture and AMD VERSAL platforms to identify research gaps, leading to a detailed report. By the second quarter, the research environment, including AMD VERSAL development kits and simulation tools, will be set up and tested. Benchmarking of the VERSAL NoC under various AI workloads will take place in the third quarter, resulting in a baseline performance report. In the final quarter, initial NoC optimization models will be developed. Risks in this phase include delays in setup and limited hardware access, which will be mitigated through the use of alternative simulation tools and close collaboration with AMD.

In Year 2, the focus shifts to exploring NoC configurations and developing optimization algorithms. In the first half of the year, simulation tools will be used to explore various NoC configurations, including routing algorithms, culminating in a detailed analysis report. The third quarter will focus on developing optimization algorithms, with initial results documented. Energy-performance trade-off analysis will begin in the fourth quarter, with an interim report detailing the findings. Potential risks include optimization algorithms not showing improvements or limited computational resources. These will be addressed through iterative refinement of algorithms and securing additional computational power through cloud services or institutional support.

Year 3 will focus on low-level data transfer optimizations and validation. The first two quarters will involve studies on data encoding and precision adjustments to optimize data transfer, with findings compiled into an optimization report. In the third quarter, the optimized NoC configurations will be integrated into a unified framework, followed by validation on AMD's VERSAL hardware in the fourth quarter. This will result in a validation report detailing the performance metrics. Risks include unexpected hardware limitations and inadequate performance improvements. These will be mitigated by collaborating with AMD for troubleshooting and iterative fine-tuning of the framework.

In Year 4, the research will culminate in a comprehensive evaluation and dissemination of results. The first two quarters will focus on evaluating the optimized NoC against the baseline, with results compiled into a performance evaluation report. In the third quarter, the final research findings will be documented, including an analysis of energy efficiency and performance improvements. The final deliverable will be a complete research report and an optimization toolkit. The last quarter will involve preparing and submitting research for publication in academic journals and conferences. Risks include delays in finalizing research and meeting publication standards, which will be mitigated by allocating buffer time for technical challenges and seeking early feedback to refine the research output.

**\* Please describe any specialist knowledge or data required to undertake your proposed research, such as language competency, technical skills or use of specialist software. If this knowledge or data is not already in place, details should be provided as to how it will be acquired over the course of the scholarship:**

This research requires a combination of specialist knowledge, technical skills, and specific software tools to optimize the Network-on-Chip (NoC) architecture within AMD VERSAL AI accelerators.

1. Technical Skills and Knowledge:

Network-on-Chip (NoC) Architecture: A deep understanding of NoC design principles, including routing algorithms, data flow management, and congestion control, is essential. I possess foundational knowledge in this area from my previous studies and research experience. During the scholarship, I will further develop this expertise through hands-on experimentation and engagement with advanced literature.

Deep Learning Accelerators and Hardware Design: Expertise in deep learning accelerators and hardware design, specifically AMD VERSAL architecture, is crucial. My background in hardware design and AI provides a strong basis. I will enhance this knowledge by studying AMD's technical documentation, research papers, and working directly with VERSAL development kits.

Data Transfer and Encoding Techniques: Understanding data transfer and encoding methods is necessary to optimize data movement within the NoC. While I am familiar with basic data management techniques, I will deepen this knowledge through literature reviews and practical experimentation focused on NoC-specific optimizations.

2. Specialist Software and Tools:

Simulation Tools (e.g., BookSim, AMD NoC Modeling Software): Proficiency in simulation tools like BookSim and AMD's NoC modeling software is key for exploring and optimizing NoC configurations. I have experience with general simulation tools and will acquire expertise in these specific tools through tutorials, documentation, and practice throughout the project.

Programming Languages (Python, C++): Strong programming skills in Python and C++ are needed for developing optimization algorithms and interfacing with simulation tools. I am already proficient in these languages, which positions me well to efficiently implement and test various NoC configurations.

Power Modeling and Profiling Tools: Energy-performance trade-off analysis requires familiarity with power modeling tools, such as those provided by AMD's VERSAL toolkit. I plan to gain proficiency in these tools by working directly within the AMD development environment and collaborating with AMD's research team.

3. Data and Access Requirements:

Access to AMD VERSAL Development Kits: Access to AMD's VERSAL development kits and the HPC research platform HACC is crucial for the real-world validation of NoC optimizations. This access is already secured through collaborations with AMD and ETH Zurich, ensuring the necessary hardware resources for this research.

Literature and Technical Documentation: Ongoing access to the latest research papers, technical documentation, and AMD's development resources is essential for keeping up with current advancements. This will guide the research direction and implementation of the NoC optimizations.

By leveraging my existing skills and systematically acquiring additional expertise through targeted learning, experimentation, and collaboration with AMD, I am well-prepared to undertake this research. This comprehensive skill set and access to essential resources will enable the successful optimization of NoC for AI accelerators, advancing both academic and industrial knowledge in the field.

**\* Please outline your plans for the dissemination and knowledge exchange of your research, including publications, conference attendance, poster presentations, reports and outreach activities. Details should also be provided as to how the impact of your research will be measured:**

Publications:

The primary dissemination of research results will be through publications in high-impact, peer-reviewed journals such as IEEE Transactions on Computers, ACM Transactions on Design Automation of Electronic Systems, and Journal of Parallel and Distributed Computing. These journals cover key areas like computer architecture, AI hardware, and embedded systems. The research will be published in phases, including initial findings on NoC benchmarking, progress in NoC configuration optimization, energy-performance trade-off analysis, and the development of a unified framework for NoC optimization in AI accelerators. Each paper will provide detailed methodology and results, contributing valuable insights into NoC optimization for the academic community.

Conference Attendance and Reporting:

Presenting at international conferences is a crucial part of the dissemination strategy. I plan to submit papers and present findings at conferences such as the IEEE/ACM International Networks-on-Chip Symposium (NoCS), International Conference on Field-Programmable Logic and Applications (FPL), and Design Automation Conference (DAC). These conferences provide not only a platform for sharing research but also opportunities for knowledge exchange with experts and industry professionals, fostering collaboration and receiving feedback to further improve the work.

Poster Presentations and Workshops:

In addition to oral presentations, poster sessions at conferences will be used to reach a broader audience. Events like the IEEE Symposium on High-Performance Computing Architecture (HPCA) will facilitate direct interaction with researchers, promoting in-depth discussions and knowledge sharing. Participation in workshops focused on AI accelerators and hardware optimization will provide additional platforms to disseminate research findings and engage with the community.

Reporting and Industry Collaboration:

Regular reports will be shared with partners, especially AMD's Versal research team, documenting research progress, methodology, and findings. These reports will foster a collaborative environment. In addition, internal seminars or webinars with industry partners like AMD will be conducted to present research outcomes, potentially influencing future product design and strategy.

Outreach Activities:

To reach a wider audience, outreach activities will include writing articles for technology blogs and contributing to open-access platforms. By simplifying research findings, the importance of NoC optimization in AI acceleration can be made accessible to a non-specialist audience. Workshops and lectures within academic institutions will also be organized to share knowledge with students and staff, encouraging the adoption of optimized NoC designs in academic projects and courses.

Impact Assessment:

The research impact will be measured through key metrics such as the number and quality of publications, citation counts, feedback received at conferences, and the adoption of NoC optimization techniques in academia and industry. Collaboration with AMD and feedback from their research team will directly measure industry relevance. Additionally, engagement in outreach activities will be tracked through views, downloads, and discussions on blog posts and open-access publications to gauge the broader societal impact.

**\* Please outline your reasons for choosing your proposed (a) academic supervisor(s) and (b) higher education institution making particular reference to how the chosen supervisor and institution**

(a) Academic Supervisor:

I have chosen Professor Shreejith Shanker as my academic supervisor due to his extensive expertise and significant contributions to the fields of computer architecture, AI hardware acceleration, and Network-on-Chip (NoC) design. His research aligns closely with my proposed study's objectives, particularly in optimizing AI accelerators and developing advanced NoC architectures. Professor Shanker's experience with cutting-edge technologies, including his work with AMD's VERSAL architecture, provides him with a deep understanding of the challenges and opportunities in AI hardware optimization. His established collaborations with industry partners like AMD offer invaluable insights and practical guidance, allowing my research to have real-world relevance. Under his supervision, I will benefit from his knowledge and mentorship, ensuring my research is both academically rigorous and practically applicable in the evolving AI hardware landscape.

(b) Higher Education Institution:

Trinity College Dublin (TCD) is the ideal institution for my research due to its strong reputation for excellence in engineering and computer science. The School of Computer Science and Statistics at TCD is renowned for its advanced research in AI, computer architecture, and embedded systems, offering a vibrant and intellectually stimulating environment. The college's access to state-of-the-art facilities, including advanced hardware platforms like AMD's VERSAL, provides the necessary resources for conducting high-impact research. Furthermore, TCD’s strategic partnerships with industry leaders, such as AMD, facilitate valuable practical engagement, enhancing the real-world impact of my study. The research community at Trinity is known for its focus on innovation and societal impact, aligning perfectly with my goal to advance NoC optimization for AI accelerators.

By choosing Professor Shreejith Shanker and Trinity College Dublin, I will have access to exceptional guidance and resources crucial for the success of my research. Their combined expertise, facilities, and industry connections will greatly enhance my ability to make meaningful contributions to the field of AI hardware optimization, ensuring that my work is both academically and industrially relevant.

**\* Please provide details of any proposed research trip(s) of more than four weeks duration which you believe will be necessary for the successful completion of your award:**

Proposed Local Research Trip:

Location: AMD Research Lab, Dublin

Duration: 6 weeks

Objective: The goal of this research trip is to collaborate with AMD's research team in Dublin and gain direct access to the VERSAL hardware platform. This access is crucial for testing and validating the NoC optimization techniques developed in my research.

Reason for the Trip: While based at Trinity College Dublin, extended access to AMD’s facilities will allow for hands-on experimentation with VERSAL devices and real-time collaboration with AMD engineers. This intensive in-person work is essential for troubleshooting and fine-tuning NoC configurations, ensuring that the research aligns with industry standards and practical requirements. The uninterrupted access over this period will facilitate a deeper validation process.

Expected Outcomes: Over the 6-week period, I will conduct experiments to measure performance metrics like latency, throughput, and energy efficiency on the AMD VERSAL hardware. These findings will be critical for refining the NoC optimization techniques and providing robust real-world validation for my research.

Contribution to the Project: This trip will significantly enhance the project's success by allowing direct, real-world testing on industry-grade hardware. Collaboration with AMD’s team will provide immediate feedback, ensuring that the NoC optimizations are both theoretically sound and practically applicable. This experience is crucial for delivering impactful research in AI hardware optimization.

This local trip is an essential part of my research plan, providing the necessary resources and expertise to ensure the successful completion of my project.