Department	School of Computing
Supervisors	CHERUBIN, Stefano
Project Title	Compiler-Level Precision Tuning

PROJECT DESCRIPTION

It is currently debated whether we are approaching or already passed the end of Moore's law. Researchers are investigating new approaches to increase system performances. The approximate computing approach aims at computing smarter rather than computing more. Several techniques have been proposed to allow a degree of error or uncertainty in the result in exchange for a faster or more efficient computation. Precision tuning is among the most popular techniques and it is based on adjusting the numeric representation of data.
In this project, you will evaluate and innovate code analysis, transformation, and optimisation algorithms to improve time-to-solution and energy efficiency of the computing systems. New computing architectures based on emerging technologies will be investigated to provide a customised system to fully exploit their support for old and emerging floating point and fixed point numeric representations.

Although precision tuning has deep roots into the field of scientific calculus, its scope is mostly limited to IEEE-754 binary32 and binary64 data formats, better known to programmers with the keywords "*float"* and *"double"*. In the last few years, the trend in acceleration of machine learning workloads pushed towards the adoption of fixed point data format and smaller floating point data types, such as binary16 and bfloat16. With more than two options available, the decision on which format is the best fit becomes more complex, thus trial & error algorithms are not very efficient. Recent works (Cattaneo 2021) demonstrated effective handling of architecture-aware mixed precision problems with multiple data types. This work, however, mostly focused on sequential execution.
Parallelism- and heterogeneity-aware programming models provide an abstraction layer over complex hardware aspects, such as data persistence, layout, and movement which can be subject to mixed precision optimisations at different levels (e.g. in the host memory, in the accelerator memory, during the data transfer).
The aims of this project are to demonstrate that automatic mixed precision technologies can be applied in parallelism-aware open standards, and to push the standardisation effort by developing an open mixed-precision programming model. To achieve these aims, theoretical and engineering efforts are required.
Detailed objectives of this project are:

- Formalisation of (computer architecture-aware) approximate computing concepts applied to mixed precision tuning

- Definition of a domain-specific language (DSL) that could map the mixed precision concepts to programmer-friendly interactions within a parallelism-aware programming model

- Enhancing the state-of-the-art in mixed precision tuning to include support for an open parallel programming model

- Analyse the effect of mixed precision tuning when applied in three different scenarios (within the accelerator, within the host memory, during the data transfer).

- Develop a system software framework to support the mixed-precision DSL

- Evaluate the framework on state-of-the-art hardware accelerators featuring mixed precision algebraic operations

Prospective applicants are encouraged to contact the Supervisor before submitting their applications. Applications should make it clear the project you are applying for and the name of the supervisors.

Academic qualifications
A first degree (at least a 2.1) ideally in Computer Science or related discipline with a good fundamental knowledge of algorithms and data structures.

English language requirement
IELTS score must be at least 6.5 (with not less than 6.0 in each of the four components). Other, equivalent qualifications will be accepted. Full details of the University's policy are available online.

Essential attributes:

- Experience of fundamental C++ programming
- Competent in concurrent and parallel systems, hardware/software co-design, and code optimisations
- Knowledge of fundamentals of computer architectures
- Good written and oral communication skills
- Strong motivation, with evidence of independent research skills relevant to the project
- Good time management

Desirable attributes:

- Experience with one or more heterogeneity-aware code acceleration paradigms (OpenCL, CUDA, SYCL, etc.)
- Experience with large C++ code bases
- Experience with MLIR and/or LLVM framework
- Track record of open source code contributions

Indicative Bibliography	Cattaneo, D., Chiari, M., Fossati, N., Cherubin, S., & Agosta, G. (2021, December). Architecture-aware precision tuning with multiple number representation systems. In 2021 58th ACM/IEEE Design Automation Conference (DAC) (pp. 673-678). Cattaneo, D., Chiari, M., Agosta, G., & Cherubin, S. (2022). *TAFFO: The compiler-based precision tuner*. SoftwareX.
Enquiries	For informal enquiries about this PhD project, please contact s.cherubin@napier.ac.uk
Web page	https://www.napier.ac.uk/research-and-innovation/research-degrees/application-process