PhD Oral Defense

Date: **3 November 2021 (Wednesday)** **Time:** **3:30pm**

Thesis Title

Efficient Application-specific Hardware Architecture for Dense Tensor Computation

Mr. HUANG Weipei (Supervisor: Prof. YAN Hong
Co-supervisor: Dr. CHEUNG C C Ray)

Abstract

A tensor is defined as a multidimensional (or multiway) array; for example, a matrix is a two-dimensional tensor. The most common approach to analyse multidimensional data is to first flatten or vectorise the data and then use well-developed matrix analysis tools. However, this approach ignores the spatial information that a tensor provides. A better approach to analyse the sensor is the tensor decomposition method. To perform the tensor computation efficiently and conveniently, it is necessary to build up a hardware architecture for tensor computation. As we aim to provide novel optimized hardware architecture design to perform tensor computations, we present a hardware architecture for singular spectrum analysis of Hankel tensors, which is a special structure tensor that is useful in signal processing. In the proposed design, in general, we have 3 major modules. To minimize BRAM usage, Hankel tensor entries are computed on the fly in higher order singular value decomposition (HOSVD). The fast tensor-matrix multiplication scheme is used to accelerate core tensor calculation. In tensor reconstruction and hankelization, a fully pipeline architecture is used to accelerate the whole process. We also presents a specific hardware architecture for tensor decomposition. Non-optimised dense tensor decomposition easily consumes a large amount of memory space. Frequent and large amounts of off-chip memory access will limit the overall performance improvement. Through computation partition and rearrangement, data movement between the FPGA and off-chip DDR memory is reduced. To reduce resource usage, an efficient and unified processing element array for a three-dimensional real tensor computation is designed. The processing element array is optimised for thin and tall tensor–matrix multiplication and two types of tensor-times-matrix chain operations.