Automatic Synthesis of Algorithms on Multi Chip/FPGA with Communication Constraints

Tomohiro Maruoka, Yukio Miyasaka, Akihiro Goda, Amir Masoud Gharehbaghi, Masahiro Fujita

The University of Tokyo

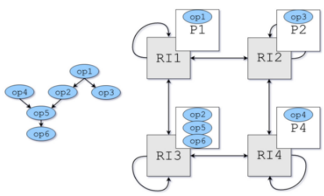
***Abstract*—Mapping of large systems/computations on multiple chips/multiple cores needs sophisticated compilation methods. In this demonstration, we present our compiler tools for multi-chip and multi-core systems that considers communication architecture and the related constraints for optimal mapping. Specifically, we demonstrate compilation methods for multi-chip connected with ring topology, and multi-core connected with mesh topology, assuming fine-grained reconfigurable cores, as well as generalization techniques for large problems size as convolutional neural networks. We will demonstrate our mappings methods starting from data-flow graphs (DFGs) and equations, specifically with applications to convolutional neural networks (CNNs) for convolution layers as well as fully connected layers.**

I. Introduction

Optimal automatic mapping of large systems/ computations on multi-chip/multi/core structures is addressed in this work. We have proposed several methods to optimally map different classes of problems considering the communication constraints. In this demonstration, we show the implementation of those methods on a couple of applications including convolutional neural networks. Following is a brief description of our compilation/mapping methods.

A. General (irregular) Computations

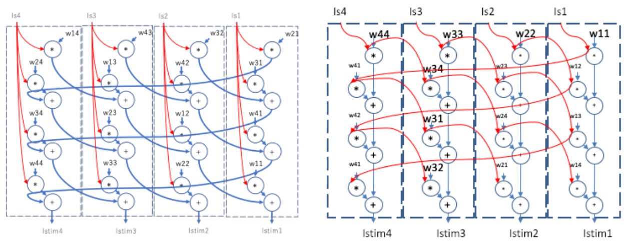
The general computations are modeled as dataflow graph and mapped to the target architecture with our ILP (integer linear programming) formulation plus the heuristics to enable optimal mapping of larger problems. However, for regular computations there are more efficient ways that will be introduced in the next sections. Our formulation and the implementation support different communication topologies, although we have mainly focused on mesh-based fined-grained reconfigurable systems [1].



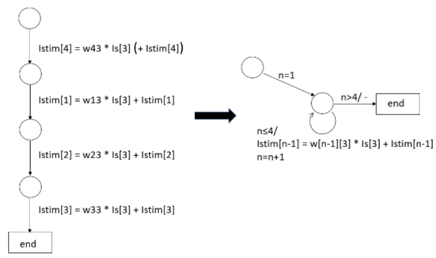
B. Regular Computations

We have proposed optimal mapping methods for regular computations, such as matrix operations or neural network computations. Our method is based on the observation that the optimal mapping of large problems can be generalized from the optimal solution of the small size problem. For example, once an optimal solution for small matrix size is obtained, the solution for very large metrices can be generalized. For regular computations, we have only considered regular and scalable communications architectures like ring and mesh.

We have proposed optimal mapping methods based on QBF (quantified Boolean formulation). For example, matrix-vector multiplication for 4x4 matrix and vector of size 4 could be done in the following ways, assuming 4 cores connected with ring topology. Note that this an essential operation in neural networks.

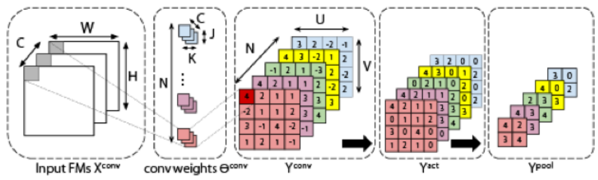


Generalization of the small size problem to large problems could be done by humans, resulting in communication and mapping constraints for larger problems. This way, larger problems can be solved and verified efficiently. However, it is desirable to automate the generalization task. We have introduced a loop-synthesis approach for this problem. Following figure shows a simple example.

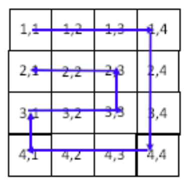


II. CNN Example

Nowadays, convolutional neural networks (CNNs) are used in many applications; hence, acceleration of the convolution operations is very important. Following is a typical CNN used in image classification.



We will demonstrate our novel way of performing convolution on a moving window by performing the multiply-accumulate operations following ring connection-like order of operations as well as moving the window on the input image to maximize the utilization of processing element in each block as well as making data transfers locally. Following picture shows an example on 4x4 window.



References

[1] A.M. Gharehbaghi, T. Maruoka and M. Fujita, "A New Reconfigurable Architecture with Applications to IoT and Mobile Computing", IFIP Internet of Things Conference (IoT 2018), Sep. 2018, Springer International Publishing, IFIP Advances in Information and Communication Technology, in press