tetracom

Technology Transfer in Computing Systems

D3.50: Individual TTP50 abstract

Project no.:	609491
Funding scheme:	Collaborative project
Start date of the project:	1st September 2013
Duration:	36 months
Work programme topic:	FP7-ICT-2013-10
Deliverable type:	Report
Deliverable reference number:	ICT-609491 / D3.50
WP and tasks contributing:	WP 3 / all
Due date:	31/07/2016
Actual submission date:	25/07/2016
Responsible Organization:	IMPERIAL
Dissemination Level:	Public
Revision:	1.0

TETRACOM D3.50: FPGA Acceleration of Stencil Computation

Tim Todman, Wayne Luk (Imperial College London), Paul Grigoras (Corerain Technologies Ltd. UK)

Stencil computation refers to a class of iterative operations to update array data with a fixed pattern, known as a stencil. Stencil computations are commonly used in (1) simulating dynamic systems, such as fluid dynamics and heat diffusion, as well as in solving Partial Differential Equations (PDEs), (2) image processing applications, and (3) deep learning algorithms such as Convolutional Neural Network (CNN). As shown in Figure 1, since neighbouring data in multiple dimensions are required for each computation, spatial locality reduces as the dimension size and the number of dimensions increase. Limited by the sparse data access patterns, performance of stencil computations is limited to 1.8 GFLOPS on a 4-core Intel i7-870 CPU for a fifth-order stencil. Propagating the stencil for 1000 time steps in 1024*1024*1024 space requires 63.4 Tera floating-point operations, and takes 10 hours to finish. The high-performance requirements limit the usage of stencil computations in scientific research and industrial development.

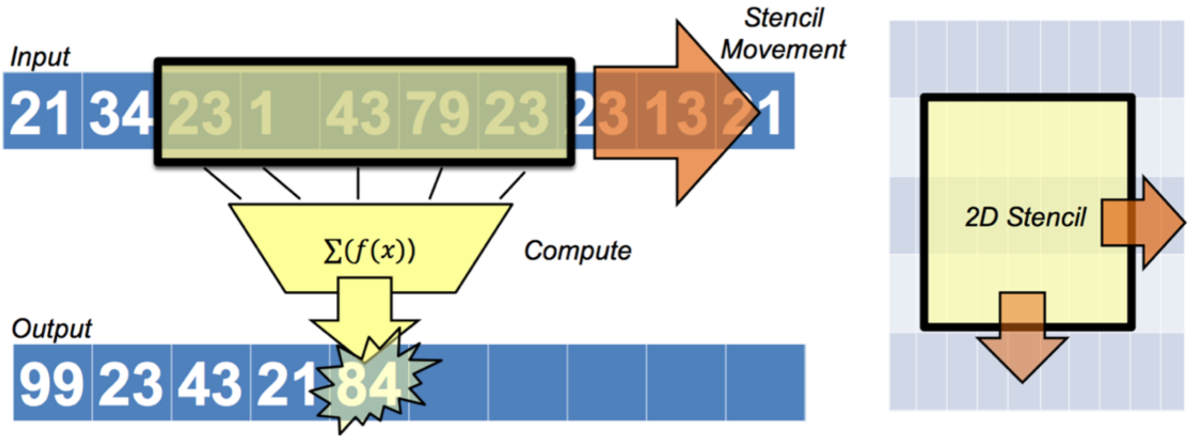

Figure 1. A 2D stencil computation example.

In this TTP, we focus on designing hardware acceleration and compiler support for stencil computation. The challenges include addressing:

- Slow computation: the limited performance of stencil computation limits the efficiency of dynamic system simulation, image processing, and deep learning applications.
- Low productivity: customized hardware architectures, while provide higher performance, are time-consuming to develop.
- Low flexibility: one customized hardware architecture cannot fit all stencil-based applications due to various data size and data type preference.

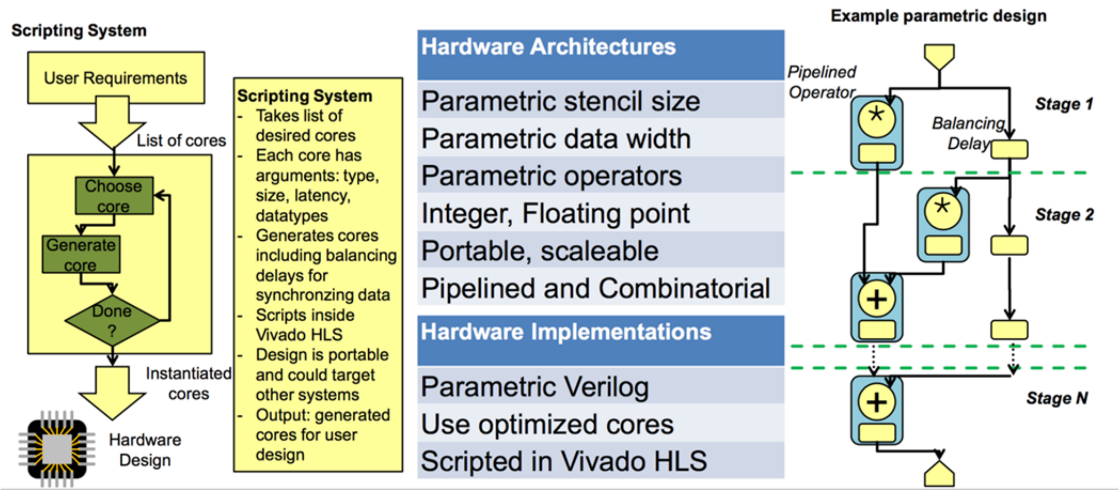

Figure 2. Tool flow overview.

Corerain Technologies, as a leading provider of real-time data analytic solutions, faces the challenge of supporting stencil computation efficiently in hardware and from high-level descriptions. The purpose of the TTP is to investigate the use of the Imperial College Custom Computing Group's customized stencil hardware architectures in developing Corerain Technologies' real-time data analytic solutions. The general goals are to improve the performance, productivity and flexibility of stencil hardware architectures.

During this TTP, a simulation platform is built to support the verification and the performance measurements of developed stencil hardware architectures. As shown in Figure 2, an automatic architecture generation system is developed to customize low-level hardware architectures based on user requirements. This improves productivity as well as flexibility for developing stencil applications. In order to generate high-performance architectures, the arithmetic operators are pipelined for maximum throughput. The supported user specification includes: stencil size, data width, data type and arithmetic operator implementation. As illustrated in the figure, the architecture generation process includes:

- Capture stencil computation cores from high-level applications.
- Combine user specifications and captured stencil computation cores, to configure hardware module properties.
- Optimize the initial hardware modules to reduce design resource usage, pipeline arithmetic operators, and to balance delays for synchronizing data.
- Generate the optimized stencil architecture in Verilog HDL, as a portable hardware module.

This turns out to be good practice to see if the academic research results could be applied in industrial environments – some extensions are made to accommodate Corerain specific requirements. In the TTP, the Custom Computing Research group has successfully integrated the automatic architecture optimization and generation process into the Corerain Streaming Insight System.