Automated design of networks of Transport-Triggered Architecture processors using Dynamic Dataflow Programs

Hervé Yviquel, Jani Boutellier, Mickaël Raulet, Emmanuel Casseau

To cite this version:

HAL Id: hal-00909325
https://hal.archives-ouvertes.fr/hal-00909325
Submitted on 26 Nov 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Automated design of networks of Transport-Triggered Architecture processors using Dynamic Dataflow Programs

H. Yviquel\textsuperscript{a}, J. Boutellier\textsuperscript{a}, M. Rault\textsuperscript{b}, E. Casseau\textsuperscript{a}

\textsuperscript{a}University of Rennes 1, IRISA, Inria, France.  
\textsuperscript{b}INSA of Rennes, IETR, France.  
\textsuperscript{c}CSE department, University of Oulu, Finland.

Abstract

Modern embedded systems show a clear trend towards the use of Multiprocessor System-on-Chip (MPSoC) architectures in order to handle the performance and power consumption constraints. However, the design and validation of dedicated MPSoCs is an extremely hard and expensive task due to their complexity. Thus, the development of automated design processes are of highest importance to satisfy the time-to-market pressure of embedded systems.

This paper proposes an automated co-design flow based on the high-level language-based approach of the Reconfigurable Video Coding framework. The designer provides the application description in the RVC-CAL dataflow language, after which the presented co-design flow automatically generates a network of heterogeneous processors that can be synthesized on FPGA chips. The synthesized processors are Very Long Instruction Word -style processors. Such a methodology permits the rapid design of a many-core signal processing system which can take advantage of all levels of parallelism.

The toolchain functionality has been demonstrated by synthesizing an MPEG-4 Simple Profile video decoder to two different FPGA boards. The decoder is realized into 18 processors that decode QCIF resolution video at 45 frames per second on a 50MHz FPGA clock frequency. The results show that the given application can take advantage of every level of parallelism.

Keywords: Co-design, RVC, Dataflow programs, MPSoC, TTA, FPGA

1. Introduction

Over the past few years, the use of multimedia applications in embedded systems has massively grown thanks to the commercial success of devices such as...
as smartphones and tablets. The inefficiency, in terms of power consumption, of General Purpose Processors (GPP) to execute multimedia applications has already been shown [10].

As a remedy to the inefficiency of GPPs, Multiprocessor System-on-Chips (MPSoCs) have been used in the embedded domain for some years already [24]. However, the design and validation of dedicated MPSoCs is an extremely hard and expensive task, even more when the MPSoCs contain heterogeneous cores. Thus, the development of automated design processes, such as the one proposed in this paper, are of highest importance for the embedded industry.

Reconfigurable Video Coding (RVC) is an MPEG initiative to provide a development framework dedicated to produce and maintain video coding tools in a modular and reusable fashion [2]. The framework is based on the dataflow programming paradigm which enables reuse and reconfiguration of the coding tools. Thanks to the explicit parallelism and modularity within dataflow, the framework is ideal for automated generation of efficient heterogeneous platforms.

This paper proposes an automated co-design flow that exploits the high-level language-based approach of the RVC framework. The designer provides the application description in the RVC-CAL dataflow language [6], after which the presented co-design flow automatically generates a network of heterogeneous Very Long Instruction Word-style processors that can be synthesized on FPGA chips and execute the application. The methodology permits the rapid design of complex many-core signal processing systems and the automation of the design flow enables easy and less error-prone development.

This paper extends preliminary work [3] in the following ways: the automatically generated code has clearly higher performance than in the earlier work; the compilation flow has been improved by the use of a better intermediate representation; the design flow is described more formally and in higher detail.

The main contributions of this paper are:

- The description of a fully automated co-design flow for instantiation of a network of heterogeneous processors from an application description based on the dataflow programming paradigm.

- The use of a new intermediate representation (IR) of the software code in order to increase the flexibility of the co-design flow. The adoption of a new IR required designing several sophisticated software transformations that are described in Section 4.2.

- A simulation infrastructure that eases the performance analysis, debugging and system integration of the proposed design. This part is explained in Section 4.3.

The paper is organized as follows. Section 2 introduces the RVC framework, the LLVM intermediate representation and the Transport-Trigger Architecture and their benefits when used in an MPSoC design flow. Related work on MPSoC design automation is presented in Section 3. Section 4 presents a precise description of the proposed design flow. Section 5 illustrates the design flow
functionality by synthesizing an MPEG-4 Simple Profile video decoder to two different FPGA boards.

2. Background

The design methodology proposed in this paper is based on several existing technologies which are briefly described below to show their benefits when used in an MPSoC design flow.

2.1. Reconfigurable Video Coding

RVC is an MPEG initiative to provide a development framework dedicated to produce and maintain video coding tools in a modular and reusable fashion [2]. The framework is based on a dataflow programming language called RVC-CAL [6] which permits the description of an application by a dataflow graph of interconnected components, called actors. The dataflow description enables reuse of the components and dynamic reconfiguration of the application. The explicit parallelism and modularity of RVC dataflow programs is ideal for automatic generation of efficient heterogeneous platforms.

RVC-CAL is a Domain-Specific Language (DSL) [22] created to help the development of signal processing systems. RVC-CAL is based on Dataflow Process Networks (DPN) [18], a special case of Kahn Process Networks (KPN) [16]. These dataflow Models of Computation (MoC) are called dynamic, because their components, called actors, can have data-dependent behavior. In other words, the behavior of an actor can depend on its input data.

This DSL is supported by the Open RVC-CAL Compiler (Orcc) [1], an open-source framework able to generate both hardware [21] and software [23] descriptions from one RVC-CAL description of an application. Orcc is based on Model-Driven Engineering (MDE) technologies [8] which speed up the design process by automating time-consuming and error-prone tasks. The Orcc project also contains the Just-in-time Adaptive Decoder Engine (Jade) [9], a software implementation of a virtual machine -based universal decoder engine.

2.2. LLVM intermediate representation

Contrary to the previous work [3], the intermediate representation (IR) used in our compilation flow is the one developed for the LLVM project (Low-Level Virtual Machine) [17]. The new IR was adopted because of its potential to carry additional information for the compiler via metadata. The potential of metadata for adaptive compilation of actors has already been shown in Jade [9]. Moreover, the LLVM IR provides the flexibility to handle the bit-accurate word lengths of the RVC-CAL programming language.

The LLVM project is an open-source compilation framework which reaches the performance of industrial compilers while maintaining modularity and reusability. As a consequence, it has been widely used in both academia and industry. LLVM provides type safety, low-level operations, flexibility and permits the
proper representation of (practically) all high-level languages. It is used during all software-targeting phases of the presented design process.

The LLVM representation was developed to be used in three different contexts: in a classical compiler as an easily analyzable and transformable intermediate representation (IR); as a Just-In-Time compiler for fast loading from an on-disk bitcode; and finally as a human-readable assembly language representation.

2.3. Transport-Triggered Architecture

The instruction processor technology used in our design flow is Transport Triggered Architecture (TTA). TTA was chosen for the following reasons:

- **Instruction-Level Parallelism**: TTA processors are able to take advantage of the only type of parallelism which is not inherent in RVC-CAL. TTA processors resemble Very Long Instruction Word processors (VLIW) in the sense that they fetch and execute multiple instructions each clock cycle. A major difference, however, is that TTA processors have only one instruction: move, which simply transfers data from a processor internal place to another. For example, one move instruction can initiate a data transfer from the output of an add execution unit of the TTA processor to one of the inputs of a multiplier execution unit. Here, the concept execution unit is used in the sense of functional unit included in most of the processors. The concept of functional unit could be confusing because it is another name for an actor in RVC.

- **Embedded processors**: TTA processor are ideal for targeting embedded systems. In [5] it is stated that direct programming of the data transports reduces the register file traffic when compared to VLIWs, but on the other hand makes the compiler design quite challenging, as it is the compiler that schedules the data transports and makes sure conflicts are avoided. Since the compiler makes these decisions at design time, the run-time system is simplified and hence there are savings on the processor gate count and energy consumption.

- **Flexible architecture**: TTA processors are extremely configurable. The designer can make the processor tiny and energy-efficient or, if needed, increase the instruction-level parallelism of the processor arbitrarily. The TTA design environment also allows the creation of custom instructions and custom execution units, which increase the processor efficiency at the cost of making the processor somewhat more application-specific. Figure 1 shows a small example of a TTA processor composed of two buses, two execution units, one register file, one load/store unit (to manage RAM accesses) and one control unit connected to the instruction memory (ROM).

- **Robust tools**: The open source TTA Co-design Environment (TCE) [7] offers a robust toolset for the design and use of TTA processors. The TCE toolset enables the design of custom processors and their realization
into VHDL files and memory images for easy FPGA synthesis. The TCE toolset is composed of a compiler which is based on LLVM [17]. It also contains a processor simulator which permit the profiling of the executed application.

3. Related work on MPSoC design flows

Park et al. classify MPSoC design approaches for signal processing systems in [20] as follows: a) the use of model-based programming languages in order to express parallelism explicitly, b) using compilation techniques to extract the parallelism from the source code and c) extending programming languages to explicitly express parallel parts of the algorithm. Our methodology consists of a mixture of a) and b): the RVC-CAL language provides data-level parallelism on the high level, whereas the TTA compiler automatically extracts the instruction-level parallelism on the low level (inside dataflow actors).

3.1. Dataflow-based approaches

Design flows from RVC applications to hardware platforms (in the sense of ASICS) have been implemented by two different tools: Openforge presented in [13] and Orc presented in [21]. The basic idea of these approaches is the direct transformation of RVC-CAL descriptions into Register Transfer Level (RTL) descriptions suitable for FPGA or ASIC synthesis. The major difference between both methodologies comes from the abstraction level of the generated code: Openforge generates low-level and optimized HDL code dedicated to a specific platform (close-to-gate RTL), whereas Orc generates high-level, portable and readable HDL code (close-to-hand-written RTL). Both approaches obtain excellent results in terms of gate count and frame rate. However, both of these methodologies suffer from a severe limitation as they are only applicable on single-rate RVC-CAL programs, i.e. actors can only read and write single tokens at once. However, [14] describes a way to handle this limitation using an automated transformation from multi-rate RVC-CAL programs to a single-rate
ones. Nevertheless, the results of both RTL-producing approaches show an explosion in the logical gate count and a significant reduction in the maximum frequency of the designs due to the complexity of the resulting code.

In [11], the authors present an architecture dedicated to the RVC methodology, composed of a set of predefined hardware components (ASIC) and an ARM processor. An actor is mapped to a hardware component if a suitable ASIC is available. Otherwise, the ARM processor executes the software description of the actor. The authors do not use any single high-level language description but resort to traditional software and hardware descriptions of the components (mostly in C and VHDL). The use of predefined ASIC components is a considerable limitation in terms of future evolution of the platform.

3.2. Compiler-based approaches

In [4], the authors present a toolset which aims at parallelizing C applications for MPSoC platform. However, the process is not automated and needs the assistance of the programmer. The method is limited to thread-level parallelism.

A manually designed TTA processor for Inverse Discrete Cosine Transform (IDCT) for a video decoder is described in [19]. As our results also show, a VLIW-like processor can easily take advantage of the instruction-level parallelism of the IDCT algorithm. The authors present a real-time framerate for a 720p sequence with a clock frequency of 200MHz, but the design only encompasses a single algorithm. Moreover, the authors of [19] do not present any results about the quantity of work of manually designing such a dedicated processor.

3.3. Language-extension approaches

In [12], the authors present a multicore TTA co-design flow for parallel programming languages OpenCL and OpenMP. Contrary to our model, the processor cores are interconnected using a shared memory and exploit mechanisms such as threads and synchronization.

A parallel programming model, called embedded Message Passing Interface (eMPI), is used in [15] to establish a complete MPSoC co-design methodology using distributed memory Network on Chips (NoC) from . However, this model is based on processes and network mechanism which results of an enormous overhead according to the fine granularity of our actors.

4. Proposed TTA-based MPSoC design flow

This section presents the automated process for generating TTA-based MPSoCs from RVC application descriptions and is organized as follows: first, the hardware design flow for generating the MPSoC HDL description is presented; then, the flow of compiling the executables for each processor is described. Finally, a description of the testing infrastructure is given.
4.1. Hardware design flow

The design approach illustrated in Figure 2 shows a direct mapping of the RVC application to a hardware network of TTA processors. Each part in the application dataflow graph is mapped to an equivalent hardware component. For example, an actor is associated to a processor and a connection between two actors is replaced by a hardware FIFO channel.

![Figure 2: Design approach](image)

Our co-design flow presented in Figure 3 is implemented around two open-source projects: Orcc and TCE. Orcc can be considered as an RVC-CAL front-end for TCE and TCE as a TTA back-end for Orcc. In practice, the design flow is decomposed to network and actor levels. The network level corresponds to the instantiations and interconnections of the components of the design (processors and FIFO channels). This step is performed by Orcc. The actor level is a full co-design flow wherein both TCE and Orcc are involved. Orcc generates a high-level description of the processors and the intermediate representation of the software code associated to each actor, then TCE uses this information to generate a complete processor design. The description of the processor enables the generation of its VHDL description using a pre-existing database of standard hardware components and the software code is compiled into executable binary code.

The processors need to be capable of executing dynamic dataflow programs. RVC programs often consist of actors that have data-dependent behavior, i.e. their execution depends on the value of their input data. To enable this, some specific execution units called stream units had to be developed. These stream units enable the communication between the concurrently executing processors over hardware FIFO channels. Figure 4 presents an example of an interconnection between two processors. Such stream units have to reproduce the behavior of RVC-CAL FIFOs, particularly their ability to give the number of available tokens and to read data without consuming it (also known as peeking).
4.2. Software design flow

The compilation flow is composed of two distinct steps as presented in Figure 5. In the first part, Orcc translates the RVC-CAL code to the intermediate representation. In latter part, performed by the TCE compiler and presented in [7], the intermediate representation is transformed to parallel assembly that is executable by a TTA processor.

The transformation from RVC-CAL code to the LLVM IR was created for this work to enable the proposed design flow. It incorporates several sophisticated transformations and optimizations that are explained below:

1. **Special FIFO operations**: Direct *FIFO read*, *FIFO peek*, *FIFO status* (acquire number of tokens) and *FIFO write* operations are instantiated to
the LLVM IR. In contrast to, e.g., memory-mapped FIFO access, these special operations allow very fast FIFO communication.

2. **Action scheduler**: In RVC-CAL, the scheduling of actions is expressed using priorities, finite-state machines, guards and constraints on the FIFO states. This transformation expresses the action scheduler in a procedural way to make it understandable by the TTA compiler.

3. **Representation properties**: A total transformation procedure had to be designed to enable making LLVM IR representations out of RVC-CAL actors by respecting properties such as Static-Single Assignment and Three-Address Code. This procedure consists of variable indexing, \( \phi \)-function addition and splitting of complex expressions to multiple primitive instructions.

4. **Correct handling of word lengths**: The RVC-CAL language allows the designer to express bit-accurately the word length of each variable and communication channel. The respective property is also found in the LLVM IR. However, when a computation is to be performed with two variables of different word lengths, the correct result must be ensured by the use of an explicit *cast* instruction, as it is done in the proposed work.

After applying these fundamental transformations, the resulting LLVM IR representation is suitable for the target-independent powerful optimizations of the LLVM compiler then the specific optimizations of the TTA compiler.

### 4.3. Proposed simulation and debugging infrastructure

Much of the difficulty of adopting MPSoC platforms is due to the following reasons:

- **Debugging** of parallel hardware is very difficult when compared to debugging of software debugging. Execution tracing of hardware blocks is very limited when compared to tracing of software executions.
Performance analysis at platform level is very difficult. Based on the performance of individual blocks, it is impossible to tell anything about the performance of the whole platform.

System integration for MPSoC is a slow and error-prone process.

Our design flow tackles these difficulties by offering a cycle-accurate simulation, using the TCE [7], which can operate at different levels:

- **Actor level**: Each processor can be tested independently from the others. The testing workbench compares automatically produced output data to reference output. A reference output is obtained by running the application on a general purpose processor (for which the C back-end of Orcc generates the software).

- **Network level**: The whole design is simulated to check the functionality of the application including the communication between the processor. This enables evaluating the performance of the application without using an FPGA board.

Moreover, our co-design flow also creates files to enable using a hardware simulator (e.g. Mentor Graphics ModelSim) to check the HDL description. The software simulator is about two hundreds time faster than the hardware one.

5. Experiments

The previous sections presented in detail the automated design flow from RVC-CAL descriptions to a hardware platform composed of a network of TTA processors.

After the generation of such a design, the designer evaluates the performance of the generated hardware components using the simulation tools. If the required performance is not reached, the designer customizes the processors that are identified as bottleneck actors.

In this section, we demonstrate the applicability of our approach using the MPEG-4 Simple Profile video decoder as a case study.

5.1. RVC-CAL description

We have used a description of an MPEG-4 Simple Profile video decoder known as MVG, which is available in the Open RVC-CAL Applications bundle at [1]. This description is composed of twenty actors (of which two are simple broadcasts), which communicate using forty FIFO channels. These actors are classified as following sub-networks: the parser is dedicated to the entropy decoding, texture is used to decompress the image texture information and motion that performs the motion compensation.

The actors are described at a fine granularity level, most of them compute one block (8x8 pixels) at a time, with very dynamic behavior.
The design generated from this description of the MPEG-4 Simple Profile video decoder is presented in Figure 6. This design is composed of 18 processors, two hardware broadcast and 40 hardware FIFOs. The performance evaluation of the generated design has been done by decoding the nine first frames of the Foreman sequence (QCIF), available on the website of Orcc.

5.2. Benchmarks

Table 1 describes three different configurations of TTA processors used during the experiments. The first one, called standard, is almost equivalent to a RISC processor: inside the TTA processor the interconnection network is composed of two buses that can provide two operands to an execution unit at each clock cycle and move the result when it is available. The two last configurations, custom and huge, define larger processors composed of several execution units and many buses able to take advantage of the instruction-level parallelism of the application (like a Very Long Instruction Word processor). The huge configuration is only used for simulation purposes to acquire the maximal performance.

The simulation results of each actor for the Standard and Huge configurations are presented in Table 2. These results represent the number of processor cycles needed to produce enough data to decode the given sequence. The simulator assumes that input data are always available during the processor execution. During a real execution, a processor may have to wait for its predecessors. This table shows the maximum performance limited by the application’s instruction-level parallelism (ILP).

Unfortunately, the Huge configuration can not be used to implement networks of TTA processors on our FPGA boards due to the limited quantity of
available logic. Consequently, we use the smaller Custom configuration for the six bottleneck actors and the Standard processor for the other ones. Table 3 shows the detailed simulation results for the six bottleneck actors of this RVC decoder. The Custom configuration is a good compromise between complexity and performance. This is confirmed by the speedup presented in table 3; it is close to the one acquired with the simulated Huge configuration.

The performance results of hardware synthesis are presented in Table 4 for two FPGA boards: Altera Stratix III (EP3SL150F1152C2) and Xilinx Virtex 6.
(XC6VLX240T). The generated designs dedicated to Xilinx and Altera boards differ only by the proprietary memory components used for RAM, ROM and FIFOs.

<table>
<thead>
<tr>
<th>Actor</th>
<th>Network</th>
<th>Standard</th>
<th>Custom</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverse quantification</td>
<td>Texture</td>
<td>70140000</td>
<td>3840000</td>
<td>1.83</td>
</tr>
<tr>
<td>Merger</td>
<td>-</td>
<td>73250000</td>
<td>2957000</td>
<td>2.48</td>
</tr>
<tr>
<td>Interpolation</td>
<td>Motion</td>
<td>73630000</td>
<td>3037000</td>
<td>2.42</td>
</tr>
<tr>
<td>Add</td>
<td>Motion</td>
<td>88820000</td>
<td>4733000</td>
<td>1.88</td>
</tr>
<tr>
<td>IDCT 2D</td>
<td>Texture</td>
<td>12110000</td>
<td>6059000</td>
<td>2.00</td>
</tr>
<tr>
<td>Frame buffer</td>
<td>Motion</td>
<td>15361000</td>
<td>7646000</td>
<td>2.01</td>
</tr>
</tbody>
</table>

Table 3: Simulation results in clock cycles for six bottleneck actors of a MPEG-4 Simple Profile decoder using two different processor configurations (Standard and Custom)

5.3. Discussion

We have demonstrated the functionality of our design flow by implementing a whole RVC-CAL MPEG-4 Part 2 Simple Profile decoder on two FPGA boards. This application is composed of actors that have very different behavior and granularity. Simulation results with the Huge configuration of the processors show two categories of actors, control and computational actors:

1. Control actors have very limited ILP, between 1 and 2 instructions per clock cycle. Their computational needs are minimal or the scheduling of
their actions is too complex to take advantage of the execution parallelism of TTA processors without software strategies like branch predication or hardware mechanisms such as a branch predictor.

2. Computational actors like the inverse discrete cosine transform (IDCT) and interpolation are the traditional bottlenecks of the MPEG-4 SP decoder. The resulting speedup for these actors, between 2.0 and 3.5, depicts their high instruction-level parallelism. They are ideal for execution on VLIW-like processors. This is the reason why it is interesting to use in this case a configuration of TTA processors providing more ILP, like the custom configuration.

The performance on the FPGA board after synthesis shows a speedup of two between the Standard configuration and Custom. In this particular application, the buffer actor remains as the bottleneck that limits the performance of the whole design.

On both FPGA boards, the use of larger processors reduces the maximum clock frequency. Indeed, the critical path of the design increases according to the complexity of the interconnection network in each processor.

In our designs, most of the processors are identical: at most two different TTA processors configurations are used to implement 18 actors. For an optimal performance / resource usage tradeoff, each actor should have a tailored processor. However, the use of identical processors is a first step towards a more generic multi-processor platform able to execute several RVC applications.

6. Conclusion

This paper presents a co-design flow for instantiating many-core systems out of a high-level application description. The many-core system is a network of heterogeneous processors based on the Transport-Trigger Architecture. The presented co-design flow allows a rapid development and evaluation process of complex signal processing applications. We have validated the method with the RVC-CAL description of an MPEG-4 Simple Profile decoder and we are able to automatically generate an FPGA implementation in a few seconds.

The work described in this paper enables future works concerning the dataflow-based design approach of signal processing MPSoCs. The programmability of the TTA processors permits the design of domain-specific platforms able to execute various applications. To reach this target, a generic interconnection model has to be defined.

The platform generation process needs to be improved to reach the performance requirements of modern signal processing systems, such as real-time HD video decoding performance. The application could be accelerated by generating a mixed platform containing some instruction processors and some hardware accelerators generated directly from the RVC-CAL code.
7. Acknowledgments

We would like to thank the following people for their contributions in the Orcc project: Matthieu Wipfiez, Antoine Lorence, Khaled Jerbi and Jérôme Gorin. We would also give special thanks to Pekka Jääskeläinen for the time and effort he took to help us with the TCE.

References


