Abstract—Architectures with multiple programmable cores are becoming more attractive for video codec applications because they can provide highly concurrent computation and support multiple video standards and a shorter time-to-market. To find an efficient SW code for the multiple core architecture for a video codec application, it is very important to easily explore the design space by generating a SW code automatically from its functional model.

We introduce Abstract Clock Synchronous Model (ACSM) for functional modeling of video codec applications. The ACSM can easily represent both parallelism and conditionals, which are common in video codec applications. By applying ACSM to an H.264 baseline decoder on single core architecture, we reduced the execution time and the number of external memory accesses by 32 % and 46 % respectively compared to traditional dataflow model.

I. Introduction

Current video codec applications require a higher performance architecture that can process more complex algorithms for higher resolution images. To find an architecture that satisfies this requirement with a limited resource, it is essential to use functional modeling that can exploit deep pipelines and parallelism and maximize the re-use of the limited resource. Furthermore, it is important to use a programmable architecture that supports multiple video formats and newly emerging standards. To improve the design productivity, both SW and HW codes should be generated automatically if possible because the time-to-market is getting shorter. Therefore, automatic SW code generation from a functional model is one of the essential technologies in system design. In this paper we focus on a functional modeling method to generate an efficient SW code automatically from the mapping result. The mapping is based on an evaluation function where the inputs are a functional model to a target architecture, and 3) generation step is based on the mapping result. We will show its efficiency by comparing it with conventional dataflow methods on a single core architecture in terms of performance, and communication bandwidth.

The rest of the paper is organized as follows. In Section II we explain ACSM, a proposed solution suitable for modeling video codec applications. In Section III ACSM is compared with previous functional modeling methods in detail. In Section IV we present several experimental results to check the efficiency of ACSM, which is followed by the conclusions in Section V.

II. Abstract Clock Synchronous Model

A. Assumptions

Fig. 1 shows the overall steps of automatic SW code generation from functional model. Three key design steps are 1) building a functional model of a target architecture, 2) mapping of the functional model to a target architecture, and 3) generation of SW code automatically from the mapping result. The mapping step is based on an evaluation function where the inputs are a functional model and a target architecture and the outputs are evaluation metrics such as performance, power and cost. Therefore, to generate efficient SW code from a functional model, the target architecture should be taken into consideration in building a functional model.

In this paper we assume that a target architecture is composed of processor (or HW) subsystems for image processing, global memory subsystems for image store, and an interconnection for communication between subsystems, as shown in Fig. 2. A processor (or HW) subsystem is composed of a processor (or HW IP), a local memory, an interconnection interface, and a local bus.
In this paper, we focus on the functional modeling step for efficient SW code generation of video codec application.

B. Abstract clock synchronous model

The Clocked Synchronous model (CSM), which has been used in designing the hardware for clocked synchronous circuits, is based on the clock synchrony hypothesis [2]: There is a global clock signal controlling the start of each computation in the system, and communication takes no time, and computation takes one clock cycle. This assumption makes it possible to describe the functionality of a circuit deterministically independent of the detailed timing of the gates in the circuit by separating each combinational logic block from others with clocked registers.

Our method is to exploit the orthogonalization between functionality and timing in the synchronous design methodology [4]. In this paper we extend the CSM to the ACSM by using an abstract clock of larger granularity that is suitable for system-level design.

Fig. 4 shows examples of (a) clocked synchronous model, (b) abstract clock synchronous model and their implementations (c-d).

In order to describe ACSM and compare it with the existing functional modeling methods, we follow the tagged-signal model introduced in [3]. Given a set of values V and a set of tags T, an event e has a tag t and a value v, i.e. e = (t, v) ∈ T × V. A signal s is a set of events. The tags are used to model time, precedence relationships, and synchronization points. The values represent the operands and results of computation.

In the ACSM, it is necessary to represent intra- and inter-iteration dependencies explicitly in order to exploit parallelism and pipeline as above mentioned. To do this, we use a set of tags T ∈ ω × ω, where ω is the set of nonnegative integers with the usual numerical order. The set of the first components of all events is totally ordered. The first component is used to model data precedence between operations across an abstract clock.
boundary, i.e. inter-iteration dependencies. The set of the second components of all events is partially ordered. The second component is used to model data precedence between functions within an abstract clock interval. The set of all event tags is partially ordered because of the second component.

Let $e_{ij}$ denote an event where the tag $t$ is $(i, j)$ and the value $v$ is $e_{ij}$. Given two events $e_{ij}$ and $e_{ilm}$, $e_{ij} < e_{ilm}$ if $(i, j) < (n, m)$. Fig 5 shows an example of precedence relationships between events in an ACSM. $F_1$ imposes a precedence constraint such that $e_{ij} < e_{ij3}$. $Delay$ imposes a precedence constraint such that $e_{ij} < e_{i+1,j2}$. But there is no precedence relationship between $e_{ij}$ and $e_{ijk}$, and $e_{ij}$ and $e_{i+1,jk}$. The partially ordered events give rise to parallel and pipelined execution of functions in an ACSM.

A data type $D$ can be extended into a data type $D_0$ by adding the special value $\bot$ to model the absence of a value at a certain tag. Absent events are used to model the outputs of unselected operations in if-then-else structures in the ACSM.

$D_{abstract clock for video codec application}$

In the ACSM, it is important to select an abstract clock suitable for explicitly representing parallelism, communication, and communication buffers for efficient design space exploration. Video decoding (and encoding) processes are basically composed of four hierarchical iterations: sub-macroblock level, macroblock level, slice level, and image level. Each of these iteration indices can be a candidate for the abstract clock. Two iteration indices of slice and image levels are too coarse to represent parallelism, and communication buffers explicitly and the decoding order of macroblocks as shown in Fig. 6 (a) is regular so that we can represent the data dependencies easily in Simulink.

$E_{Basic components and firing rules of ACSM}$

We decided to use Simulink [5] for a video codec application because it provides simulation and modeling environment of discrete-time systems that are sufficient to build an ACSM. It also includes Real-Time Workshop (RTW) [6], which generates a C code automatically from a Simulink model.

Fig. 7 shows basic components of ACSM that are expressed easily in Simulink.

- **Block (process):** A block, as shown in Fig. 7 (a), maps $n$ input events on $m$ output events: $(o_1, ... , o_m) = F_0(i_1, ... , i_n)$. It corresponds to a S-function or pre-defined block with inherent sample rate in Simulink.

- **Delay:** A delay, as shown in Fig. 7 (b), represents that its output is delayed from its input by $k$ abstract clock cycles. It corresponds to discrete delay in Simulink.

- **Arc:** An arc carries events from one output port of a block or a delay to one or more input ports of one or more blocks and/or delays as shown in Fig. 7 (c). It corresponds to connecting line in Simulink.

We also defined two kinds of subsystems that are composed of blocks, delays, arcs, and other subsystems.

- **If-action subsystem (IAS):** An IAS, as shown in Fig. 7 (d), represents an if-then-else structure. An IAS is enabled when its control input port, which is connected to an if/else block, has a present event, i.e. not absent. If an IAS is not enabled, its output ports have absent events. All output ports must be connected to a merge block and only one of them can have a present event at a time. It corresponds to "If-action subsystem" in Simulink.

- **For-iterator subsystem (FIS):** A FIS, as shown in Fig. 7 (e), represents a for-loop structure. It is used to describe sequential or parallel repeated executions of blocks where the number of repetitions is known. It corresponds to "For-iterator subsystem" in Simulink. A FIS usually includes Demux and Max. A Demux divides an event into several (sub-)events. A Max integrates several (sub-)events into an event.

$F_{An example of ACSM}$

Fig. 8 shows a simplified ACSM in Simulink for the H.264.
baseline profile decoder. It includes two paths. A path consists of macroblock VLD (MB VLD), 8x8 sub-macroblock inverse quantization (8x8 IQ), and inverse quantization (8x8 IT) to compute a residual image from a video bit stream. The other path consists of MB VLD, spatial compensation (SC) or motion compensation (MC) from the current frame or previous frames. If four neighbor 4x4 sub-macroblocks, e.g., 0, 1, 2, 3 blocks in Fig. 6(b), have the same motion vector, it is possible to fetch less image data from previous frames (8x8 IF in IAS1) and manipulate the image data more efficiently (8x8 MC in IAS2) compared to 4x4 sub- macroblock based motion compensation (4x4 IF and 4x4 MC in IAS3) by the elimination of common computation and communication. The ACSM consists of 83 S-functions, 286 arcs, 21 computation and communication. Therefore, pre-fetchs. It also requires context switching to deal with blocked processes. To reduce the overhead of context switching, it is necessary to increase the granularity of process.

In the Kahn process networks [7], concurrent processes communicate through one-way FIFO channels with unbounded capacity. This means that writes to the channel always succeed immediately, while reads block until there is sufficient data in the channel to satisfy them. In particular, a process cannot test an input channel for the availability of data and then branch conditionally.

Fig. 9 (b) shows a KPN example corresponding to an ACSM model as shown in Fig. 9 (b). A process in a KPN model has a computation code mixed with a communication code. Therefore, the KPN doesn’t support explicit and predictable communication that is required to efficiently use burst data transfers and data pre-fetchs. It also requires context switching to deal with blocked processes. To reduce the overhead of context switching, it is necessary to increase the granularity of process.

A. Kahn Process networks

Fig. 9. Functional model examples of different styles.

IV. Comparison with previous modeling styles

In order to obtain an efficient SW code from a functional model, it is important to select an appropriate functional modeling style according to the property of an application domain. In the following subsections, we will compare ACSM with the most popular previous modeling styles in respect of building functional models of video codec applications. The capabilities and the weakness of these models will be analyzed with regards to the requirements stated in section I.
D. Synchronous model

Synchronous model (SM) [10] used in Esterel [11] and Lustre [12] is based on the perfect synchrony hypothesis assuming that the reaction to each set of the inputs is considered to be instantaneous. In the SM, a process is fired when it has at least one event on its input ports. When a process is fired, it can test absent events on its input ports and produce absent events on its output ports.

The synchronous assumption simplifies system specification and verification. However, difficulties arise especially for video codec applications if the target architecture is a distributed multi-core system, because it is very expensive to maintain a global clock for testing and producing absent events over a distributed system. The fully synchronous implementation based on time-triggered architecture [13] must be conservative, forcing the global clock to run as slow as the slowest computation and communication process. Therefore, SM is not suitable for video codec applications, which have large variations in computation and communication.

E. Analysis: comparing with ACSM

The ACSM overcomes all the restriction of the above mentioned models while providing the same scheduling facilities. In data-driven models of KPN and SDF, the absent token is not defined. Therefore they have some difficulties in expressing an explicit if-then-else structure. In the BDF model the token overflow problem comes from the definition of the absent token without a global clock. However, the ACSM uses an abstract clock for the well-defined absent token. In the ACSM, an event on an arc is updated at every abstract clock tick, so there is no token overflow problem.

Although the dataflow model has difficulties in representing conditionals, it can express valid schedules of a multi-rate algorithm with only a model. In a SDF model as shown in Fig. 10 (a), a schedule that requires minimal buffers is F0F0F1F2F2F0F1F2F2 and another schedule with loop structures is (3F0)(2F1(2F2)). Contrary to the dataflow model, the ACSM can express only one schedule with a model. Fig. 10 (b) shows an ACSM example that expresses a schedule that is (3F0)(2F1(2F2)). However, it is possible to find other schedules by loop transformation techniques such as loop unrolling and loop split.

In the SM, there is no restriction on producing and testing absent events. It can cause a causality problem and requires a global clock to solve the problem. However, the ACSM allows only restricted absent events to express if-then-else structures. In other words, it is sufficient for a distributed multi-core system to replace absent events of unselected IAS outputs with present events only if an if-then-else structure is implemented over different processors.

V. Experiments

We performed several experiments to check the efficiency of the ACSM compared to the dataflow model, assuming that target architecture consists of a processor subsystem for image processing and an external global memory for image storage, as shown in Fig. 11, where a processor subsystem consists of a Tensilica Xtensa processor [14] with a default configuration, a local memory, and a DMA with a memory interface. We assumed that image fetch processes are executed in the DMA module.

The target application is an H.264 baseline decoder and its input bitstream is the Foreman sequence of QCIF format, which was encoded withQP=28 and IntraPeriod=5 when all MC block sizes were enabled.

Fig. 12 shows our experimental procedure. In the modeling step, we made three Simulink models from H.264 decoder reference C code as shown in Table I. In the Simulink model equivalent to a SDF model, an if-then-else structure is expressed by using a MUX block as shown in Fig. 9 (c) and the processing block size is 4x4. In Simulink model equivalent to an ACSM, an if-then-else structure is expressed by an if/else block and IASs and the processing block size is 4x4. In the optimized ACSM, the processing block sizes are either 4x4 or 8x8. For 8x8 block, an additional IAS is necessary as shown in Fig. 8. We limited the block size of the SDF model to 4x4 because the additional IAS causes redundant computation as explained in IV-B.

Then, we used Real-Time Workshop (RTW) to generate SW codes from the three different Simulink models. In the final step, we used Xtensa gdb and gprof to measure the execution cycles and the external memory access of the SW codes. We also added the reference C code to the test sets.

Fig. 13 shows the execution time and the number of external memory accesses for the four different configurations. According to the experimental results, the ACSM reduces both the execution time and the number of external memory accesses by 29 % and 22 % respectively compared to those of the SDF model. Furthermore, the optimized ACSM reduces them by 32 % and 46 % respectively compared to those of the SDF model.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDF</td>
<td>SDF model as shown in Fig. 9 (c)</td>
</tr>
<tr>
<td>ACSM</td>
<td>Proposed ACSM model as shown in Fig. 9 (a)</td>
</tr>
<tr>
<td>Optimized ACSM</td>
<td>Proposed ACSM model with 8x8 block size as shown in Fig. 8</td>
</tr>
<tr>
<td>Handed C code</td>
<td>Reference C code</td>
</tr>
</tbody>
</table>

Fig. 13 shows the execution time and the number of external memory accesses for the four different configurations. According to the experimental results, the ACSM reduces both the execution time and the number of external memory accesses by 29 % and 22 % respectively compared to those of the SDF model. Furthermore, the optimized ACSM reduces them by 32 % and 46 % respectively compared to those of the SDF model.
According to the experimental results, it is necessary to express conditionals such as if-then-else structures explicitly for building a functional model of a video codec application. The optimized ACSM code from RTW requires 41% more execution time compared to the handed code because it includes many redundant copy operations.

Even if a SDF model with embedded controls can remove redundant computation, it cannot remove the external memory access as explained in section IV. It also limits the design space exploration due to increasing the granularity of blocks.

For the BDF model of the H.264 baseline decoder, 138 SWITCHes were required to resolve the token overflow problem mentioned in section IV. The number of the SWITCHes in the BDF model is larger than that of S-functions in the ACSM model, which means that although the BDF supports conditionals, it is not suitable in building a functional model for a video codec application because the BDF functional models get too complex.

VI. Conclusions

In this paper, we explained a functional modeling method for generating an efficient SW code from a functional model for video codec applications. It is based on ACSM, which is an extension of the clocked synchronous model by employing the macroblock index as an abstract clock. The ACSM can express conditionals easily by allowing absent events with the global abstract clock. It can also express parallelism and pipeline easily by partially ordered intra- and inter-dependencies. Therefore, the ACSM is suitable for functional modeling of video codec applications that require both parallelism and conditionals.

Experimental results with Simulink and RTW showed that the SW code generated from an ACSM of H.264 decoder is improved by up to 32% and 46% compared to its SDF model in terms of the number of execution cycle and the number of external memory access respectively. We found that a BDF model of H.264 decoder requires many additional SWITCHes to express if-then-else structures. So the BDF model gets more complex. Therefore, the ACSM is more effective in building functional model for video codec applications because emerging standards such as H.264 require complex data-dependent operations.

For a H.264 baseline decoder, the SW code generated with the ACSM by RTW requires longer execution time and larger buffer memories compared to those of the handed code. We are in the process of developing a more efficient SW generation tool. Furthermore, we will extend this methodology for multi-processor systems because the current version of RTW can generate only a SW code for single-processor systems.

Fig. 13. Relative execution times and external memory accesses of different configurations.

Acknowledgements

The authors are grateful to Xavier Guerin, Paul Amblard, Frédéric Pétrot and Nacer-Eddine Zergainoh from TIMA laboratory for their inputs on this paper. This work was supported by ITSoC Project and Brain Korea 21 Program.

References