Abstract

Innovation in the field of embedded systems, and more broadly in cyber-physical systems, increasingly relies on software. The productivity gain in software development cannot keep up with the demand for software despite the increasing adoption of Model-Driven Development (MDD). In this context, we believe that major productivity and quality improvements are still ahead of us through better programming languages and environments. CPAL, the Cyber-Physical Action Language, is a contribution in that direction with the objective to speed-up the development of embedded systems with dependability constraints. The objective of this paper is to present and illustrate the use-cases of the high-level abstractions offered to the developer in CPAL with respect to real-time scheduling, introspection mechanisms, native support of Finite State Machines (FSMs), abstracting the hardware and decoupling functional concerns from non-functional concerns.

Categories and Subject Descriptors D.2.11 [Software Engineering]: Software Architectures

Keywords Cyber-Physical Systems, Embedded systems, Model-driven development, control applications, dependability

1. Introduction and related work

Context of the work. Innovation crucially relies on software today and software is actually disrupting complete fields of the economy and the industry. Embedded systems, and Cyber-Physical Systems at large, are no exception whatever their application domains: vehicles, medical devices, home appliances, factory automation, power distribution, Internet of Things (IoT), etc. If the amount of software is growing fast, the productivity gains in software development are much slower. In this landscape, Model-Driven Development is certainly a key technology but, in our view, programming languages still lack the high-level abstractions and automation features (“state the what, not the how”) that would make them more productive.

The objective of CPAL is to offer a solution to fill this gap for the domain of embedded systems. CPAL is a high-level interpreted language supporting a model-driven development flow for the development of timing-predictable embedded systems. CPAL can be used in a simulation mode early in the early design phases, or, in real-time mode, to execute programs on top of an OS or on a bare-metal platform. CPAL serves to describe both the functional behavior of the functions, that is their code, as well as the functional architecture of the system (i.e., the set of functions, how they are activated, and the data flows among the functions). CPAL is however not a full-fledged Architecture Description Language able to describe complex system architectures, like AADL [8] or Chariot [14] are, but a programming language that can possibly be used to develop software components from systems described in an ADL. As discussed in this paper, in addition to the functional description, the developer can also express timing-related non-functional concerns.

CPAL has been used in a number of academic and industrial case-studies, such as the development of a smart parachute add-on component for UAVs [7] and an AUTOSAR compliant engine function [17], proposing an answer to the Thales FMTV challenge [3] and developing a simulation model of an automotive middleware on top of Ethernet [15]. CPAL has also been used since 2013 as the supporting language to teach Model-Driven Engineering for embedded systems at the University of Luxembourg. The CPAL documentation [11], graphical editor and the execution engine for Windows, Linux and Raspberry Pi platforms are freely available from http://www.designcps.com. To experiment with CPAL without any installation, a “playground” is available on the web site along with code examples illustrating the language.

Why a new domain-specific programming language? The design and development of embedded systems with dependability constraints necessitates the adoption of many good practices throughout the entire development cycle. The need to develop a new language originates from our analysis that none of the programming languages we were aware of possessed all the features we think needed to combine productivity and code correctness:
• General-purpose programming languages do not offer all the right abstractions for today’s real-time embedded systems: scheduling periodic activities, time as a first class citizen, safe inter-process communication, native support for finite-state machines, high-level interfaces to I/Os, support for timing and formal verification, etc. As an illustration, just consider how cumbersome it is to program in C a periodic activity on a real-time OS, let alone programming a set of tasks with real-time constraints sharing resources and synchronizing constraints. We note that an alternative to developing a new language is to adapt, by domain-specific extensions and restrictions, an existing language. In the field of embedded systems, this is what has been with the mbeddr framework [19] which is based on C. Such an approach has limits about what can be done since the underlying language dictates many things, but possesses advantages for instance in terms of tool-set availability and ease of adoption by the programmers.

• Synchronous languages (Lustre [6], Signal[4], etc) are meant for critical applications but they impose many constraints and a programming style that is very specific and does not suit everyone. The initial learning curve is steep too. In addition, many of the most appealing synchronous languages, such as Prelude [13] or Giotto [10], are actually only Architecture Description Languages (ADL) which means that the algorithms have to be coded in an other language. An important benefit of synchronous languages is that formal proof support, both in the time domain and value domain, is usually available, whereas CPAL currently only offers schedulability analysis to check for possible deadline misses (see [1]) and simulation.

• Correctness starts with simplicity: simplicity of the language constructs, but also of the runtime environment. The CPAL language is Turing complete but relatively small in the number of constructs and abstractions it defines, which facilitates the learning and teaching of CPAL. Although code-generation is currently being investigated, CPAL is for the time being an interpreted language whose execution engine is in the range of 10,000 lines of code (see §2.2). Especially when the execution engine runs directly on the hardware, without an OS, the amount of code involved is several order of magnitude less than with compiled code which is of course makes it easier to verify the code correctness.

The reader is referred to [12] and [2] for a more thorough comparison with other programming languages and, more generally, programming paradigms, especially those belonging to the realm of synchronous approaches.

2. High-level abstractions for the domain of critical embedded systems

The main objective of CPAL is to provide high-level abstractions suited to express domain-specific patterns of behaviors and enforce the good programming practices of the domain of embedded systems. In this Section, after an introduction to CPAL and its development environment, we focus on features related to time, introspection and interactions with the hardware. Another requirement in the design of CPAL has been to facilitate the writing of correct code. For instance, there is no untyped data, no pointers, no dynamic memory after initialization, only simple control-flow constructs, a loop over construct ensuring a bounded number of loop iterations, communication channels with LIFO or FIFO semantics, etc. Some statements which are known to be error-prone are also forbidden in CPAL, such as testing the equality of two floating-point variables.

Figure 1. CPAL code implementing a servo tester as a periodic process. The logic is described as a FSM whose transitions between states are triggered by a Boolean condition that evaluates to true, or after a certain time spent in a continuous manner in a state. The access to hardware is abstracted through the use of global variables mapped onto the I/O ports.
2.1 Basics of CPAL: processes and automata

We here first introduce the basic features of CPAL through the example of a servo tester whose code is shown in Figure 1. At the core of CPAL is the concept of process which is a built-in type for a recurrent activity of the system. Processes can be seen as functions that are activated periodically, or when certain activation conditions are met. The process is first defined (line 1 to 28 in Figure 1), like a class is in object-oriented languages, then it is instantiated with certain activation conditions and global variables - possibly mapped to I/O ports as in the servo tester example - as input and output parameters (line 34 to 36 in Figure 1). Processes can have arguments in input and output; here the servo tester takes two input variables and outputs the servo signal. Embedded in each process is a Finite-State Machine (FSM), possibly reduced to a single state, that helps to describe the logic of the process in a correct and non-ambiguous manner.

A process has a memory in the sense that, when it is activated, it resumes in the state in which it was at the end of the previous activation but the code of the state is not executed at this point. A transition out of this state is first taken if any can be, and only then the code of the current state of the FSM is executed. A specific situation is the very first execution of the process where it is assumed that a dummy transition doing nothing leads to the process initial state. After the execution of the state code, the execution of the process is then finished and the CPU becomes available for another process that is pending execution, the choice of which by the execution engine depends on the scheduling parameters (see §2.4). The transition-first semantics allows for faster response to external events and the transitions can execute code in CPAL.

The CPAL language is fully textual but views are produced out of the code to inform the developers, and possibly other stakeholders, about certain facets or behaviors of the program. Currently, the CPAL-Editor building on Graphviz is able to generate views of the FSMs in the processes, as in Figure 2 for the servo tester example, a Gantt chart of the process executions over time (see Figure 8) and a view of the functional architecture, that is the set of processes making up an application and the flows of data among them. In a first phase of the project, we developed the prototype of a graphical editor to specify the FSMs but decided to re-allocate the engineering efforts to other areas more critical at this stage of the project such as the timing correctness of the interpretation engine. If a graphical editor could certainly be helpful to some users, we believe that for users with prior programming experience, the loss of productivity of writing textual code is limited; the important thing in our view being that the FSMs can be visualized.

2.2 Execution environments and execution modes

During the development, the interpreter is executed from within the CPAL-Editor. As soon as a change in the source files is detected, the CPAL-Editor automatically parses, analyzes and executes the code in background to generate the views out of it. CPAL can also be executed in a co-simulation environment. For the development of control systems, the execution in Matlab/Simulink as an S-function is supported (see [16]). CPAL is also the simulation language to program high-level protocol layers in the RTaW-Pegase embedded network simulator from RTaW. The CPAL code that is deployed on a target is either launched from a shell or uploaded in memory and executed immediately after the startup on bare-metal platforms.

There are two execution modes in CPAL: the simulation mode and the real-time mode. In the latter mode, the timing characteristics of the processes are respected (e.g., periods) and the programs can access I/O ports. Depending on the platform, the execution engine runs on top of an OS (on Linux, Windows, Mac OS X and Raspbian) or on the bare hardware (Freescale FRDM-K64F). In simulation mode, the code is executed as fast as possible, thus not in real-time, and do not have access to the I/O ports. Through timing and scheduling annotations (see Section 3), it is however possible to enforce a more timing-realistic behavior by simulating execution times and possibly execution jitters. An interesting feature of CPAL in terms of productivity is that the simulation code can be re-used without changes in real-time mode on a target. All platforms possess both modes, except the FRDM-K64F board, which does not support the simulation mode. This platform however offers the best real-time predictability in real-time mode since it does not have the interferences from an OS. The reader is referred to [11] for more information about the execution modes and platforms capabilities.

2.3 Working with time durations

Performing actions at certain time points in time and being able to express time quantities is a common need in embed-
ded programming. To measure and manipulate time quantities, CPAL possesses the native \textit{time64} data type. The time units available are s, ms, us, ns, ps and Hz (Hertz). The latter unit is a natural unit in many applications to express the rate of activation of activities. It can however only be used for assignment and defining the periods of the processes, and not for doing arithmetic on time quantities like the other time units since the meaning of such operations would be unclear.

CPAL provides addition, subtraction and modulo operators between \textit{time64} values and multiplication and division between \textit{time64} and \textit{uint64}. The use of \textit{time64} quantities, and the \texttt{sleep()} and \texttt{time64.time()} functions with the usual semantics, is illustrated in Figure 3.

2.4 Introspection mechanisms and dynamic reconfiguration

The ability to implement adaptive behaviors requires that the activities of an application can learn about themselves during the execution. For this purpose, CPAL offers introspection features enabling to query at run-time the pid, period, offset, priority, deadline and activation time of the current and previous instance of any process. The example in Figure 4 shows a process detecting at run-time that its start-of-execution jitter is very high, which is detrimental to the quality of control in many systems. The threshold condition depends here on the actual rate of execution of the process, which eases the portability and genericity of the code.

Once an abnormal timing behavior has been detected, appropriate measures can be taken. First, a change in the functioning mode of the application can be triggered, and the FSMs embedded in the processes are well suited to describe applications with different modes and the transitions between the modes. Another possibility is to adapt the characteristics of the tasks and the quality of service offered by the run-time environment. In CPAL, some characteristics of the system can be reconfigured at run-time using (non-functional) annotations to the code (see Section 3). For instance, it is possible to change the rate of activation of a process, increase its priority if Non-Preemptive Fixed-Priority scheduling (NPFP) is used, or reduce its deadline if Non-Preemptive Earliest-Deadline First (NPEDF) is used.

Another useful design pattern to detect unwanted conditions occurring at run-time is to have a process dedicated to the monitoring of the other processes or a subset of them. Whenever needed, this supervision process can take appropriate measures, such as initiating an error recovery strategy or a mode change.

2.5 Interacting with the hardware

A domain-specific language for embedded systems should offer a good support to the programmer for sensing and actuating operations. In CPAL, global variables can be mapped to I/O ports to abstract the hardware from the programmer point of view. By default, these I/O mapped variables used by a process are updated in reading when the process starts to execute, and in writing at the end of its execution. Sometimes, as in the example of Figure 5, it is however needed to update I/O mapped variables during the execution of the process, this can be done by an explicit call to \texttt{IO.sync()}. The synchronous languages we are aware of do not provide such a possibility and impose a single write operation at the end of execution.

The mapping between the global variables and the I/O ports is in the version of CPAL available at the time of writing (version 1.15) achieved through a naming convention (e.g. \texttt{pin_gpio_c10_in} in Figure 1). To ease the porta-
Figure 5. Update of the I/O ports through the use of the IO.sync() function. I/O ports are transparently updated when the process arguments are evaluated and upon the completion of the process execution.

Figure 6. Timing annotations used in simulation to capture the execution times observed at run-time. The Matlab/Simulink port of CPAL [16] enables to re-inject these delays in the control loops for the design of timing-sensitive controllers.

3. A clear separation between functional and non-functional concerns

The correctness of a real-time application is twofold. First, the results of the computation must be correct which is called the correctness in the value domain. Second, the timing behavior of the application must be correct too, this is the correctness in the time domain. Typically, the latter involves that deadlines must be met and jitters occurring at run-time (i.e., process start-of-execution and end-of-execution jitters) must be kept within acceptable bounds. Timing is an important non-functional concern but there are other dimensions of importance in CPS such as power consumption, security and safety. We discuss hereafter the timing concerns which have been the focus of CPAL until now.

In CPAL, functional and non-functional concerns are decoupled. The non-functional concerns are expressed in annotations, placed inside or outside the code, as the programmer decides it, but anyway well identified as not belonging to the code itself. In simulation mode (see §2.2), the annotations are used for instance to experiment scheduling strategies or reflect the processing power of different execution platforms. The same code can be re-used in real-time execution mode at a later stage, but then, many annotations will be ignored. Typically this is the case for the annotations describing the execution time of the code, that are used in simulation to obtain a more timing-realistic behavior as shown in the Gantt diagram of Figure 8.

3.1 Simulating execution times

In simulation mode, it is possible to take into account the time it takes to execute the code of a process through a @cpal:time annotation. The execution time can be expressed at the level of a state (see lines 4-6 in Figure 6), the level of a block (called named block in CPAL, see lines 11-15 in Figure 6) or globally. In the later case, the annotation is defined at the global scope for each instance of a process and holds whatever the state in which this instance of the process is. The execution time can be static or dynamic, it can depend on the value of a condition (as on line 24 in Figure 6) or it can be the result of a computation. Annotations are implemented as blocks of code that are evaluated immediately before the activation of the processes.

State-level and block-level annotations obey the same logic but cannot be used interchangeably as they may result in different timing behavior. With state-level annotations, all code statements belonging to the state are executed at the time of the entry in the state; the time being only incremented by the execution time upon the exit of the state. Whereas with block-level annotations, the time is incremented upon the exit of the block, which means that the remaining instructions coming afterward will be executed at the new value of time. Block-level annotations can be used in transitions too and permits a fine-grained timing characterization of the code. In practice, the execution times in the processes can be derived by monitoring the execution on the target using the --stats option of the interpreter, or, for more accuracy, with an oscilloscope. The later solution must
be used on a bare-metal platform. A set of execution time annotations can be defined for each execution platform.

### 3.2 Simulating release jitters

In many systems, due for instance to limited time accuracy or runtime overhead, tasks will not be released and ready to execute exactly when they are scheduled to be. A certain delay, called release jitter, may happen and vary from a task instance to the next. This can be modeled in CPAL using a `@cpal:time` annotation as illustrated in the code snippet below.

Although formally the release jitter is something different from the start-of-execution jitter seen in §4 that is mainly caused by the execution of higher priority processes, the release jitter annotation can be used to aggregate all the latencies created by the other processes at early stages of the development process when the other processes are not precisely known.

```plaintext
1 const time64: min_jitter = 0ms;
2 const time64: max_jitter = 10ms;
3 4 processdef Process_With_Release_Jitter() {
5     /* .. */
6 }
7 8 process Process_With_Release_Jitter: p1[100ms] {
9     @cpal:time:p1{
10         p1.jitter = time64.rand_uniform(min_jitter, max_jitter);
11     }
12 }
```

### 3.3 Specifying scheduling policies

The default CPAL scheduling policy is FIFO: processes are executed in the order of their release. Although FIFO performs rather poorly in terms of meeting deadlines (see [1]), especially when there are tasks with large execution times, it possesses a unique property among all non-idling policies: the execution order of the processes is the same whatever the execution time of the code, and thus whatever the execution platform. This property called *event-order determinism* ensures that simulated code will behave identically to the deployed code with respect to execution order of the “significant” events of the program (i.e., sensing and actuating points). CPAL provides also the Non-Preemptive Earliest Deadline First (NPEDF) and Non-Preemptive Fixed Priority scheduling policies (NPFP). Both policies require to set one parameter per process, respectively the deadline and the priority. These parameters can be defined with a `@cpal:time` annotation.

Currently the scheduling policy is specified to the execution engine through the `--policy` command-line option, we plan to implement a dedicated annotation so that the policy itself can be more easily configured, and dynamically changed depending on the functioning mode of the application.

### 4. Ongoing work

CPAL has been designed with two objectives related to non-functional concerns which are driving the ongoing developments:

1. “*State the what, not the how*”: the idea is that the developer should state the permissible behavior of the program, and a system synthesis step involving both analysis and optimization generates executable artifacts that guarantee the requirements to be met. The feasibility of this idea with respect to scheduling has been assessed in [2, 18] but other dimensions such as power consumption and dependability remains to be studied. A question of particular interest in the targeted application domain of CPAL is how an application can be made sufficiently robust to respect a given safety level (e.g., SIL2) by introducing at the synthesis step fault-tolerant mechanisms such as for instance redundant executions or watchdog mechanisms.

2. “*Timing-equivalent behavior between simulation and execution*”: a program is executed and verified in simulation mode on a workstation and the same program later deployed on an embedded board executes with a timing-equivalent run-time behavior. Timing-equivalence will have different meanings depending on the applications but this property would significant ease the design of latency-sensitive applications such as control systems. How to achieve such a property is a question currently
investigated in a PhD thesis in our group at the University of Luxembourg.

CPAL is currently being extended to multicore and distributed systems. The approach taken is to have one interpreter per computational resource with synchronization points between the interpreters when needed to ensure that all interpreters have a consistent view of the system. This approach has been validated on the simulation of relatively simple multicore systems and networks but its efficiency on embedded targets and more complex simulation models remains to be ascertained. Although model interpretation brings benefits [17] and is suited for many applications, it does not cover all use-cases mainly because of the slowdown with respect to compiled code. We are considering several approaches ranging from allowing "opaque" binary functions to be called from CPAL code, like in the MAUDE language [9], to partial or complete code generation. Finally, we are currently extending CPAL in the domain of communication to better support IoT applications.

Acknowledgments

CPAL has been a team effort since its development started in 2012. The authors wish to thank all contributors to CPAL, especially Lionel Havet and Julien Rische from RTaW who have been the main developers of the interpretation engine.

References


