Simulink, simulation, code generation and tasks

Marco Di Natale Associate Professor, Scuola S. Anna - Italy, UTRC Visiting Fellow

Simulink model

Many things...

- A network of blocks
- Workspace variables
- Type declarations (bus objects)
- Matlab code (.m)
- Possibly external code
- Simulator configuration
- Code generation configuration



Simulink semantics and flow preservation

- The system is a network of functional blocks b_j Blocks can be:
 - regular (Dataflow) blocks or
 - Stateflow (state machine) blocks.



- Dataflow blocks can be of
 type continuous, discrete or triggered
- Discrete blocks are activated at periodic time instants and process input signals, sampled at periodic time instants producing a set of periodic output signals and the state updates.

Simulation flow



- Model is compiled before simulation
 - Rates are computed, values/types are propagated …
- Initialization stage at the beginning of the simulation
 - Init of simulation structure (entrate, uscite, stati ecc.)
 - Init matrices and variables
- Compute next simulation time for variable rate
- Compute outputs at next major time step
- Update discrete states at next major time step
- Integration of systems with continuous state
- Terminate and cleanup

Functional representation: SR Simulink modeling

• Simulink system = networks of blocks

$$S = \{b_1, b_2, \dots, b_n\}$$

- Blocks can be Regular or Stateflow blocks
- Regular blocks can be Continuous of Discrete type.
- All types operate on (right)continuous type signals.
- Blocks may have a state S_i or may be stateless.



Functional representation: SR Simulink modeling

- Continuous-type blocks are defined by a set of differential equations
- Discrete-type blocks are activated at events e_j belonging to a periodic sequence with 0 offset and period T_j
- When a model generates code, continuous blocks must be implemented by a fixed-step solver, with period *Tb*
- *T_b* (*base period*) must be a divisor of any other *T_j* in the system



Functional representation: SR Simulink modeling

• At each e_j the block computes its out update and state update functions, updating the values on its output signals $S_j^{new}, \overline{o}_j = f(S_j, \overline{i}_j)$



Simulink models (execution order - feedthrough)

Stateflow (or state machine) blocks react to a set of events $e_{j,v}$, derived from signals (generated at each rising or falling edge).

As such, events belong to a set of discrete time bases kT_{iv}



Simulink models (execution order - feedthrough)

$$b_i \rightarrow b_j$$

If two blocks b_i and b_j are in an input-output relationship (one of the outputs of b_i is the input of b_j), and b_j is of type feedthrough), then

$$b_i \rightarrow b_j$$

In case b_i is not of type feedthrough, then the link has a delay,

$$b_i \xrightarrow{-1} b_j$$

Semantics options

• Signals are persistent (Simulink)



• Signals are not persistent



• Algebraic loops (causal loops without delays) result in a fixed point and lack of compositionality

Semantics and Compositionality

- Semantics problem: systems compositions do not behave according to the semantics of the components
 - The problem is typical of SR semantics when there are causal cycles: existence of a fixed point solution cannot be guaranteed (i.e. the system may be ill-defined)
 - When multirate blocks are in a causal loop the composition is always not feasible



Simulink models (execution order - feedthrough)



May be a problem in a code implementation with (scheduling) delays



Simulink models (execution order - feedthrough)



Let $b_i(k)$ represent the *k*-th occurrence of bi (belonging to the set $\bigcup_{v} kT_{i,v}$ if a state machine block, or kT_i if a standard block), a sequence of activation times $a_i(k)$ is associated to b_i .

 $n_i(t)$ is the number of times b_i is activated before or at t.

In case $bi \rightarrow bj$, if $i_j(k)$ is the input of the *k*-th occurrence of b_j , then this input is equal to the output of the last occurrence of b_i that is no later than the *k*-th occurrence of b_i

 $i_i(k) = o_i(m)$; where $m = n_i(a_i(k))$

If *the link has a delay*, then the previous output value is read, $i_j(k) = o_i(m - 1)$:

From model to code

- The code generation framework follows the general rule set of the simulation engine and must produce an implementation with the same behavior (preserving the semantics).
- Goal 1: preservation of the synchronous assumption:
- The reaction (the outputs and the next state) of the system must be computed before the next event in the system.
- Goal 2: (looser property, equivalent to untimed simulation), called *flow preservation*.
- The execution of the system must guarantee
- $i_j(k) = o_i(m)$; where $m = n_i(a_j(k))$ (1)

Simulink models (feedthrough)



Simulink models (SR)



Simulink models (not feedthrough)



Example of generated code



Output

Simulink models (feedthrough)



Simulation of models

- Simulation of Multirate models
 - order all blocks based upon their topological dependencies
 - The RTW tool (meant for a single processor implementation) generates a total order based on the partial order imposed by the feedthrough semantics
 - In reality, there are many such total orders that satisfy the dependencies!
 - Other choices are possible
 - In multiprocessor implementations this can be leveraged to optimize the implementation
 - Then, for simulation, virtual time is initialized at zero
 - The simulator scans the precedence list in order and execute all the blocks for which the value of the virtual time is an integer multiple of the period of their inputs
 - Simulated execution means computing the block output and then computing the new state

From Models to implementation

• Simulink case

elist

Purpose	List simulation methods in the order in which they are executed during a simulation
Syntax	elist m:mid [tid: TID] elist < gcs s: sid> [mth] [tid: TID] elist < gcb sid:bid> [mth] [tid: TID]
Description	elist m:mid lists the methods invoked by the system or nonvirtual subsystem method corresponding to the method id mid (see the where command for information on method IDs), e.g., sldebug @19): elist n:19 RootSystem.Outputs 'vdp' [tid=0] : 0:0 Integrator.Outputs 'x1' [tid=0] 0:1 Outport.Outputs 'Out1' [tid=0] 0:2 Integrator.Outputs 'x2' [tid=0]
	Block id Method Block Task id

From Simulink models to update functions



From Simulink models to update functions



Task implementations (of multirate systems)

- In multitask implementations, the run-time execution of the model is performed by running the code in the context of a set of threads under the control of a prioritybasedreal-time operating system (RTOS).
- The function-to-task mapping consists of a relation between a block update function (or a set of them in the case of an Stateflow block) and a task, and a static scheduling (execution order) of the function code inside the task.
- The *i*-th task is denoted as τ_{i} .
- $\mathcal{M}(f_j, k, i)$ indicates that the step (update) function f_j of block b_j is executed as the *k*-th segment of code in the context of *i*.
- Legal task mappings must guarantee the block
 execution constraints

Flow preservation

• The implementation of a SR model should preserve its semantics so to retain the validation and verification results.



Simulation: zero logical execution time and zero logical communication time



From Models to implementation

• Simulink case (single task implementation)

Mode	Single-Rate	Multi-Rate
SingleTasking	Allowed	Allowed
MultiTasking	Disallowed	Allowed
Auto	Allowed	Allowed
	(defaults to SingleTasking)	(defaults to MultiTasking)

Table 2-3: Permitted Solver Modes forReal-Time Workshop Embedded Coder Targeted Models

From Models to implementation

• Simulink case (single task implementation)



Implementation of models

- Implementation runs in real-time (code implementing the blocks behavior has finite execution time)
- Generation of code: Singletask implementation



From Models to implementation

• Simulink case (single task implementation)

```
rt_OneStep()
{
   Check for interrupt overflow or other error
   Enable "rt_OneStep" (timer) interrupt
   ModelStep-- Time step combines output,logging,update
}
```

Single-rate rt_OneStep is designed to execute *model_*step within a single clock period. To enforce this timing constraint, rt_OneStep maintains and checks a timer overrun flag.

Model implementation: single task



System base cycle = time to execute the longest system reaction

- The implementation can
 use
 - Single task executing at the base rate of the system
 - A set of concurrent tasks, with typically one task for each execution rate, and possibly more.



From Models to implementation

• Multitask implementation

}

```
rt OneStep()
{
  Check for base-rate interrupt overflow
  Enable "rt OneStep" interrupt
  Determine which rates need to run this time step
  ModelStep(tid=0) --base-rate time step
  For i=1:NumTasks -- iterate over sub-rate tasks
    Check for sub-rate interrupt overflow
    If (sub-rate task i is scheduled)
      ModelStep(tid=i) --sub-rate time step
    EndIf
  EndFor
```

Generation of code: multitask mode

- The RTW code generator assigns each block a task identifier (tid) based on its sample rate.
- The blocks with the fastest sample rates are executed by the task with the highest priority, the next slowest blocks are executed by a task with the next lower priority, and so on (Rate Monotonic)



Nondeterminism in time and value

- However, this can lead to the violation of the zero-execution time semantics of the model (without delays) and even to inconsistent state of the communication buffer in the case of
 - low rate (priority) blocks driving high rate (priority) blocks.
 - high rate (priority) blocks driving low rate (priority) blocks.



Model implementation: multi-task



Model implementation: multi-task



Adding determinism: RT blocks

- Solution: Rate Transition blocks
 - added buffer space and added latency/delay
 - relax the scheduling problem by allowing to drop the feedthrough precedence constraint
- The mechanism can only be implemented if the rates of the blocks are harmonic (one multiple of the other)
 - Otherwise, it is possible to make a transition to the gcd of the blocks' periods, at the price of additional space and delay
- RT Blocks are only for intracore communication
RT blocks: High rate/priority to low rate/priority



RT blocks: Low rate/priority to high rate/priority



Limitations in the use of RT blocks (1)



Tradeoffs and design cycles

- RT blocks are **not** a functional entity
 - but an implementation device
- RT Blocks are only required
 - because of the selection of the RM scheduling policy in slow to fast transitions
 - because of the possibility of preemption in both cases
- In both cases, time determinism (of communication) is obtained at the price of additional memory
- In the case of slow to fast transitions, the RT block also adds a delay equal to the period of the slowest block
 - This is only because of the Rate monotonic scheduling
 - Added delays decrease the performance of controls

RT blocks: Low rate/priority to high rate/priority



Or... letting the sender have a priority higher than the receiver

Against RM and more difficult to schedule





- Consistency issues in the 1-1 communication between blocks with different rates may happen:
 - When blocks are executed in concurrent tasks (activated at different rates or by asynchronous events)
 - When a reader may preempt a writer while updating the communication variables (reader with higher priority than writer)
 - When the writer can preempt the reader while it is reading the communication variables (writer with higher priority).
 - Necessary condition for data inconsistency is the possibility of preemption reader — writer or writer — reader
- Also, we may want to enforce time determinism (flow preservation)

Consistency issues



- Also, a relaxed form of time determinism may be required
 - Input coherency: when inputs are coming from multiple blocks, we want to read inputs produced by instances activated by the same event

Guaranteeing data consistency

- Demonstrate impossibility of preemption between readers and writers
 - Appropriate scheduling of blocks into tasks, priority assignment, activation offsets and using worst-case response time analysis
- Avoid preemption between readers and writers
 - Disabling preemption among tasks (blocks) (RES_SCHEDULER in OSEK)
- Allow preemption and protect communication variables
 - Protect all the critical sections by
 - Disabling interrupts
 - Using (immediate) priority ceiling (semaphores/OSEK resources)
 - Problem: need to protect each use of a communication variable. Advantage (does not require extra buffer memory, but only the additional memory of the protection mechanism)
 - Lock-free/Wait-free communication: multiple buffers with protected copy instructions:
 - Typically w. interrupt disabling or kernel-level code
 - Problem: requires additional buffer memory (How much?). Advantage: it is possible to cluster the write/read operations at the end/beginning of a task, with limited change to existing code.
- The best policy may be a mix of all the previous, depending on the timing contraints of the application and on the communication configuration.

Demonstrating impossibility of preemption

- Assign priorities and offsets and use timing analysis to guarantee absence of preemption
- Input data:
 - Mapping of functional blocks into tasks
 - Order of functional blocks inside tasks
 - Worst-case execution time of blocks (tasks)
 - Priorities assigned to tasks
 - Task periods
 - (relative) Offset in the activation of periodic tasks (o_{wr} = minimum offset between writer and reader activations, O_{wr} maximum offset between the activations)
- Computed data
 - Worst case response time of tasks/blocks (considering interferences and preemptions) R_r for the writer R_w for the reader
- Two cases:
 - Priority writer > priority reader
 - Priority reader > priority writer

Absence of preemption/High to low priority

• Condition for avoiding preemption writer→reader (no assumptions about relative rates of reader/writer)



Absence of preemption/Low to high priority

 Condition guaranteeing absence of preemption or reader to writer (reader→writer)



Absence of preemption/Low to high priority

These conditions are ultimately used by the Rate Transition block
 mechanisms !!



Avoiding preemption

• Disabling preemption



The response time of the high priority block/task is affected, need to check real-time properties

- What buffering mechanisms are needed for the general case ?
 - Event-driven activation
 - One-to-many communication



- What buffering mechanisms are needed for the general case ?
 - Stream preservation (requirement)
 - Event-driven activation
 - One to many communication





- The time the buffer index is assigned (activation of the block) may differ significantly from the time when the index is actually used (at running time) because of scheduling delays
 - Support from the OS is needed for assigning indexes at block activation times



- Many issues
 - Defining efficient mechanisms for assigning indexes to the writers and the readers (if they are executed at kernel level)
 - Sizing the communication buffers (given the system characteristics, how many buffers are needed?)



Model implementation: multi-task

• Efficient but issues with data integrity and time determinism





Buffer sizing methods

Two main methods

- preventing concurrent accesses by computing an upper bound for the maximum number of buffers that can be used at any given time by reader tasks. This number depends on the maximum number of reader instances that can be active at any time.
- Temporal concurrency control. The size of the buffer can be computed by upper bounding the number of times the writer can produce new values, while a given data item is considered valid by at least one reader.

Bounding the maximum number of reader instances

 the size is equal to the maximum number N of reader task instances that can be active at any time (the number of reader tasks if d≤T), plus two more buffers: one for the latest written data and one for use by the writer [Chen97] (no additional information is available, and no delays on the links).



Temporal concurrency control

• Based on the concept of datum lifetime. The writer must not overwrite a buffer until the datum stored in it is still valid for some reader.



Combination

- A combination of the temporal concurrency control and the bounded number of readers approaches can be used to obtain a tighter sizing of the buffer.
- Reader tasks are partitioned into two groups: fast and slow readers. The buffer bound for the fast readers leverages the lifetime-based bound of temporal concurrency control, and the size bound for the slow ones leverages information on the maximum number of reader instances that can be active at any time. Overall, the space requirements are reduced.

Combination

- Readers of τ_{wi} are sorted by increasing lifetime (I_i \le I_{i+1}). The bound

$$NB_{w_i,j} = \left\lceil \frac{l_j}{T_w} \right\rceil$$

- Applies to readers with lifetime $\leq l_i$ (fast readers).
- Once j is chosen, the bound is



Wait free solution with flow preservation (slight modification to Chen&Burns protocol)

```
Data: BUFFER [1,...,NB]; NB: Num of buffers
                                                                       1 ReaderLP activation();
  Data: READINGLP [1,...,n_{ln}]; n_{ln}: Num of lower priority readers
                                                                       2 begin
  Data: READINGHP [1, ..., n_{hn}]; n_{hn}: Num of higher priority readers
                                                                       3
                                                                             constant id; – Each lower priority reader has its unique id;
  Data: PREVIOUS, LATEST
                                                                             integer ridx;
                                                                       4
                                                                             READINGLP [id]=0;
1 GetBuf();
                                                                       5
                                                                             ridx = LATEST:
2 begin
                                                                       6
                                                                             CAS(READINGLP [id],0,ridx);
       bool InUse [1,...,NB];
                                                                       7
3
                                                                             ridx = READINGLP [id];
       for i=1 to NB do InUse [i]=false;
                                                                       8
4
                                                                       9 end
       InUse[LATEST]=true;
5
       for i=1 to n_{lp} do
                                                                      10 ReaderHP_activation();
6
           i = READINGLP [i];
7
                                                                      11 begin
           if i !=0 then InUse [j]=true;
8
                                                                             constant id; - Each higher priority reader has its unique id;
                                                                      12
       end
                                                                             integer ridx;
0
                                                                      13
       for i=1 to n_{hp} do
                                                                             READINGHP [id]=0;
10
                                                                      14
           i = READINGHP[i];
                                                                             ridx = PREVIOUS;
11
                                                                      15
           if j !=0 then InUse [j]=true;
                                                                             CAS(READINGHP [id],0,ridx);
12
                                                                      16
                                                                             ridx = READINGHP [id];
       end
13
                                                                      17
       i=1:
                                                                      18 end
14
                                  Runtime part
       while InUse [i] do ++i;
15
                                                                      19 Reader_runtime();
       return i;
16
                                                                      20 begin
                                  (in the task
17 end
                                                                             Read data from BUFFER [ridx]:
                                                                      21
                                                                      22 end
18 Writer activation();
                                  code)
19 begin
                                                                                            Activation-time part
       integer widx, i;
20
       widx = GetBuf();
21
                                                                                            (supported by the OS
22
       PREVIOUS = LATEST:
      LATEST = widx;
23
                                                                                            or hooks)
       for i=1 to n<sub>hp</sub> do CAS(READINGHP [i], 0, PREVIOUS);
24
       for i=1 to n_{lp} do CAS(READINGLP [i], 0, LATEST);
25
26 end
27 Writer_runtime();
28 begin
       Write data into BUFFER [widx];
29
30 end
```

Multicore adaption of RT block

High rate to low rate communication: with explicit intercore activation signal and with synchronized activation with offsets



Multicore adaption of RT block

High rate to low rate communication: with explicit intercore activation signal and with synchronized activation with offsets



Model-based design: a functional view

- Advantages of model-based design
 - Possibility of advance verification of correctness of (control) algorithms
- Possible approaches
 - 1. The model is developed considering the implementation and the platform limitations
 - include from the start considerations about the implementation (tasking model and HW)
 - PROS (apparent)
 - use knowledge about the platform to steer the design towards a feasible solution (in reality, this is often a trial-and-error manual process)
 - CONS (true)
 - the model depends on the platform (updates/changes on the platform create opportunities or more often issues that need to be solved by changing the model)
 - Analysis is more difficult, absence of layers makes isolating errors and causes of errors more difficult
 - the process is rarely guided by sound theory (how good is the platform selection and mapping solution?)
 - Added elements (Rate-transition blocks) introduce delays
 - 2. The model is developed as a "pure functional" model according to a formally defined semantics, irrespective of the possible implementation
 - The model is then refined and matched to a possible implementation platform. Analysis tools check feasibility of an implementation that refines the functional semantics and suggest options when no implementation is feasible (more ...)

Model-based design: a functional view

- Advantages of model-based design starting from a purely functional model
 - Possibility of advance verification of correctness of (control) algorithms
 - Irrespective of implementation
 - This allows an easier retargeting of the function to a different platform if and when needed
 - The functional design does not depend on the platform
 - The verification of the functional design can be perfomed by domain experts (control engineers) without knowledge of SW or HW implementation issues
- Necessary assets to leverage these advantages ...
 - Capability of defining rules for the correct refinement of a functional model into an implementation model on a given platform
 - Capability of supporting design iterations to understand the tradeoffs and the changes that are required when a given functional model cannot be refined (mapped) on a given platform

Model-based development flow

• Platform-based design



Platform-dependent modeling: an example



PBD and RTOS/platform



Design/Scheduling trade-offs

However ...

- if the communication is fast-to-slow and the slow block completes before the next instance of the fast writer, the RT block is not required
- if the communication is from slow to fast, it is possible to selectively
 preserve the precedence order (giving higher priority to the slow block)
 at the expense of schedulability
 - Two tasks at the same rate, one high priority, the other low priority



An approach

Required steps

- Definition of the network of functional blocks with feedthrough dependencies
- Definition of the synchronous sets
- Priority assignment and mapping into tasks
- Definition of the block • order inside tasks



Conclusions

- Schedulability theory and worst-case timing analysis …
 - From the run-time domain to the design domain (already happening)
 - From the analysis domain to the optimization (synthesis) domain
 - Complemented by sensitivity analysis and uncertainty evaluation
- However ...
 - Typical deadline analysis is not enough!
 - Tasks and messages are not the starting point (semantics preservation issues from functional models to tasking models)
 - Worst case analysis needs to be complemented
 - Mixed domains (time-triggered / event-triggered)

Thank you!

