Why Real-Time Computing?

Giorgio Buttazzo

Today’s control systems are implemented in software and often require the concurrent execution of several periodic and aperiodic activities interacting through a set of shared memory buffers storing global data. Most of the times, controllers are designed assuming an ideal and predictable behavior of the execution environment; in other cases, a constant delay is only considered in the loop to take computer interferences into account. Unfortunately, reality is more complex, and the operating system may introduce unpleasant surprises in control applications. This paper presents the potential problems that may arise in a computer controlled system with real-time characteristics and proposes a set of methods for ensuring predictability and guaranteeing that the system behaves as predicted in the design.

Most of today’s control systems are implemented in software and often require the concurrent execution of several periodic and aperiodic activities interacting through a set of shared resources (typically consisting of memory buffers storing global data). In a modular design, each sensory acquisition activity is normally associated with a periodic task that gets data from the sensory board and puts them into a shared memory buffer. Control activities are also implemented as periodic tasks that read data from buffers, compute the output values of the control variables and put the results in other buffers, used by device drivers to actually drive the actuators. In some cases, especially when the system is based on a hierarchial architecture, control and acquisition activities are required to run at different rates.

Figure 1 illustrates a typical control application consisting of six tasks: boxes represent shared resources and disks represent tasks (acquisition, processing, control, and output tasks are denoted by A, C, P, and O, respectively). Each layer of this control hierarchy effectively decomposes an input task into simpler subtasks executed at lower levels.

The top-level input command is the goal, which is successively decomposed into subgoals, or subtasks, at each hierarchical level, until at the lowest level, output signals drive the actuators.

Sensory data enter this hierarchy at the bottom and are filtered through a series of sensory-processing and pattern recognition modules arranged in a hierarchical structure.

Each module processes the incoming sensory information, applying filtering techniques, extracting features, computing parameters, and recognizing patterns.

Sensory information that is relevant to control is extracted and sent as feedback to the control unit at the same level; the remaining partially processed data is then passed to the next higher level for further processing. As a result, feedback enters this hierarchy at every level. At the lowest level, the feedback is almost unprocessed and hence is fast-acting with very short delays, while at higher levels feedback passes through more and more stages and hence is more sophisticated but slower. To correctly implement such a hierarchical control structure and execute all tasks in a predictable fashion, several issues have to be taken into account:

- First of all, the operating system should provide a proper support for periodic tasks execution and include an internal mechanism to automatically activate a task at the beginning of each period.
- The scheduling algorithm should maximize the processor utilization and enable the user to analyze the system behavior to verify the schedulability of the task set (that is, to check whether each task is able to complete within its deadline).
- The kernel should also detect possible deadline misses due to execution overruns, to allow the application to react with proper recovery actions.
- When tasks cooperate through mutually exclusive resources protected by classical semaphores, the system is prone to priority inversion phenomena [21], which can cause unbounded delays on tasks execution due to complex bloc-

---

G. Buttazzo, Scuola Superiore Sant'Anna, Italy
king interferences. To prevent such a problem, the kernel should provide suitable concurrency control protocols to correctly access shared resources.

- Non blocking communication mechanisms are also essential for exchanging information among periodic tasks running at different rates. In fact, a blocking factor could cause a fast task to slow down, synchronizing with the rate of the slower task. This problem can be solved by using special asynchronous buffers with override semantics and non-consumable messages.

- When control applications perform intensive I/O operations, interrupt handling routines can create high and unpredictable interference on control activities, causing extra delays and jitter that could even jeopardize system stability. To limit the interference of interrupt drivers, suitable periodic service mechanisms need to be adopted in the kernel to protect the execution of critical tasks.

Unfortunately, these problems are often ignored and digital controllers are normally designed by neglecting the characteristics of the machine running the application, so implicitly assuming an ideal and predictable behavior of the execution environment. For simple control systems running on a computer with abundant resources, such a design practice does not cause any problem in the control system. However, when the application has a high degree of concurrency or it runs on an embedded platform with limited resources, the consequences become visible and may significantly degrade system performance. For some critical conditions, the non-ideal behavior of the execution environment can even cause the instability of the controlled system [7]. For example, an unpredictable delay introduced in a task controlling an inverted pendulum could cause the pole to cross the critical angle and fall down.

The objective of this paper is to illustrate the possible problems that can be introduced in a control application by the operating system and present a set of solutions that can be easily adopted to make the control system more predictable.

**Supporting periodic tasks**

Periodic activities represent the major computational load in a real-time control system. For example activities such as actuator control, signal acquisition, filtering, sensory data processing, action planning, and monitoring, need to be executed with a frequency derived from the application requirements.

A periodic task is modeled as an infinite sequence of instances or jobs, that cyclically execute the same code on different data. Each task \( \tau_i \) is characterized by a worst case execution time \( C_i \), a period \( T_i \), and a relative deadline \( D_i \). The \( k \)-th job of task \( i \) is denoted as \( \tau_{i,k} \) (where \( k = 1, 2, \ldots \)). Assuming that all the tasks are simultaneously activated at time \( t=0 \), the release time \( r_{i,k} \) of job \( \tau_{i,k} \) (that is, the time at which the task becomes ready for execution for the \( k \)-th time) is given by \( r_{i,k}=(k-1)T_i \). Then, the job executes for a time \( c_{i,k} \leq C_i \) and must complete its execution by an absolute deadline equal to \( d_{i,k}=r_{i,k}+D_i \). The finishing time of a job \( \tau_{i,k} \) is denoted as \( f_{i,k} \) and the worst-case response time of a task is defined as the maximum finishing time relative to the release time, that is,

\[
R_i = \max_k (f_{i,k} - r_{i,k}).
\]

To correctly support periodic task execution and relieve the user from defining its own timers, a real-time operating system should have the following features:

- First of all, it should include a programming interface for creating periodic tasks, allowing the user to specify the typical periodic task parameters (that is, the computation time \( C_i \), the period \( T_i \), and the relative deadline \( D_i \)).

- Then, it should enforce periodicity by a kernel mechanism for automatically activating a task at the beginning of each period.

- Finally, it should perform a runtime check to verify that each job executes for no more than its worst-case execution time \( C_i \) and completes no later than its absolute deadline \( d_{i,k} \). An overrun exception should be generated if \( c_{i,k} > C_i \), and a deadline miss exception if \( f_{i,k} > d_{i,k} \).

Figure 2 illustrates typical system calls used for creating and activating a periodic task. A sample periodic control task is also shown.

The `task_endcycle()` system call is needed to notify the kernel that the job finished its cycle and needs to be suspended until the beginning of the next period. All suspended jobs have to be enqueued in an IDLE queue, ordered by release times. The kernel is responsible for checking the system clock to wake up all the jobs that reached the beginning of their next period.

Figure 3 illustrates the typical state transition diagram for a periodic task. Notice that, in order to check for execution overruns, the kernel must have a precise time accounting mechanism to keep track of the actual job execution. Unfortunately, only very few
operating systems (e.g., Shark [22]) include such a feature.

Task scheduling

The test to verify the schedulability of the task set heavily depends on the scheduling algorithm and on the task model.

In 1973, Liu and Layland [17] analyzed two scheduling algorithms for handling periodic tasks: Rate-Monotonic (RM), which assigns each task a fixed priority directly proportional to its activation rate, and Earliest Deadline First (EDF), which assigns tasks a dynamic priority inversely proportional to the current absolute deadline, so the task with the highest priority is the one with the earliest absolute deadline. Given a set of \( n \) periodic tasks with relative deadlines equal to periods \( (D_i = T_i) \), Liu and Layland were able to relate the feasibility of the task set with the processor utilization, defined as

\[
U = \sum_{i=1}^{n} \frac{C_i}{T_i}.
\]

They proved that, if tasks do not block on synchronous operations,
- A set of \( n \) periodic real-time tasks is schedulable by RM if \( U \leq n(2^{1/n} - 1) \).
- A set of \( n \) periodic real-time tasks is schedulable by EDF if and only if \( U \leq 1 \).

Note that the RM bound provides only a sufficient condition for guaranteeing the schedulability, hence a task set could be schedulable by RM even for higher utilizations less than or equal to 1. Clearly, no algorithm can schedule a task set if \( U > 1 \). The RM bound decreases with \( n \) and tends to the following limit value:

\[
\lim_{n \to \infty} n(2^{1/n} - 1) = \ln 2 \approx 0.69.
\]

In spite of its lower utilization bound, the RM algorithm is widely used in real-time applications, mainly for its simplicity. In fact, being a static scheduling algorithm, it can easily be implemented on top of commercial operating systems, using a set of fixed priority levels, whereas EDF requires an explicit kernel support for managing absolute deadlines.

However, EDF is superior to RM under several aspects and allows the application to fully exploit the available resources, reaching up to 100% of the available processing time. Dertouzos [9] showed that EDF is optimal among all on line algorithms, meaning that if a task set is not schedulable by EDF, then no algorithm can ever produce a feasible schedule for that task set. When the utilization is less than one, the residual fraction of time can be efficiently used to handle aperiodic requests activated by external events. In addition, compared with RM, EDF generates a lower number of preemptions, thus causing less runtime overhead due to context switches. A detailed comparison between RM and EDF can be found in [6].

A more precise schedulability test for RM can be performed using the Hyperbolic Bound [4], according to which a periodic task set is schedulable by RM if

\[
\prod_{i=1}^{n} \left( \frac{C_i}{T_i} + 1 \right) \leq 2. \tag{1}
\]

The schedulability analysis for tasks with relative deadlines less than or equal to periods can also be performed, but it is more complex for both algorithms. The Deadline Monotonic (DM) algorithm, proposed by Leung and Whitehead [16], extends RM to handle tasks with relative deadlines less than or equal to periods.

According to DM, at each instant the processor is assigned to the task with the shortest relative deadline. A necessary and sufficient test to verify the schedulability of a periodic task set was independently proposed several authors [2, 13, 15]. It consists in computing the worst-case response time \( R_i \) of each task and verifying whether it is less than or equal to its relative deadline. The worst-case response time is derived by summing its computation time and the interference caused by tasks with higher priority:

\[
R_i = C_i + \sum_{k \in hpi(i)} \left\lfloor \frac{R_k}{T_k} \right\rfloor C_k. \tag{2}
\]

where \( hpi(i) \) denotes the set of tasks having priority higher than task \( i \) and

\[
\left\lceil x \right\rceil
\]

denotes the ceiling of a rational number, i.e., the smaller integer greater than or equal to \( x \). The equation above can be solved by an iterative approach, starting with \( R_i(0) = C_i \) and terminating when \( R_i(s) = R_i(s-1) \). If \( R_i(s) > D_i \) for some task, then the task set cannot be feasibly scheduled by DM. Under EDF, the schedulability analysis for periodic task sets with deadlines less than periods is based on the Processor Demand Criterion [3].

According to this method, a task set is schedulable by EDF if and only if, \( U < 1 \) and, for every interval of length \( L > 0 \) the overall computational demand is no greater than the available processing time, that is, if and only if

\[
\text{Processor Demand Criterion} \tag{3}
\]
Shared resource protocols

When tasks interact through shared memory buffers, the use of classical synchronization mechanisms, such as semaphores or monitors, can cause a phenomenon known as priority inversion [21]. It refers to the case in which a high priority task is blocked by a low priority task for an unbounded interval of time. Such a situation can create serious problems in critical control systems, since it can jeopardize system stability.

Priority Inheritance Protocol

An elegant solution to the priority inversion phenomenon caused by mutual exclusion is offered by the Priority Inheritance Protocol (PIP) [21]. Here, the problem is solved by dynamically modifying the priorities of tasks that cause a blocking condition. In particular, when a task \( \tau_i \) blocks on a shared resource, it transmits its priority to the task \( \tau_0 \) that is holding the resource. In this way, \( \tau_0 \) will execute its critical section with the priority of task \( \tau_i \). In general, \( \tau_i \) inherits the highest priority among the tasks it blocks. Moreover, priority inheritance is transitive, thus if task \( \tau_i \) blocks \( \tau_0 \), which in turn blocks \( \tau_0 \), then \( \tau_i \) will inherit the priority of \( \tau_0 \) through \( \tau_0 \).

Figure 5 illustrates how the schedule shown in Figure 4 is changed when resources are accessed using the Priority Inheritance Protocol. At time \( t_3 \), when \( \tau_2 \) blocks on the semaphore (direct blocking), \( \tau_1 \) inherits the maximum priority among the tasks blocked on that resource, thus it continues the execution of its critical section at the priority of \( \tau_2 \). Under these conditions, at time \( t_4 \), task \( \tau_2 \) is not able to preempt \( \tau_1 \), hence it blocks until the resource is released (push-through blocking). In other words, although \( \tau_2 \) has a nominal priority greater than \( \tau_3 \), it cannot interfere during the critical section, because \( \tau_3 \) inherited the priority of \( \tau_1 \). At time \( t_4 \), \( \tau_2 \) exits its critical section, releases the semaphore and recovers its nominal priority.

As a consequence, \( \tau_1 \) can proceed until its completion, which occurs at time \( t_6 \). Only then, \( \tau_2 \) can start executing. The Priority Inheritance Protocol has the following property [21]: Given a task \( \tau \), if \( \lambda \) is the number of lower priority tasks sharing a resource with a task with priority higher than or equal to \( \tau \) and \( \sigma \) is the number of semaphores that could block \( \tau \), then \( \tau \) can be blocked for at most the duration of \( \min(\lambda, \sigma) \) critical sections. Although the Priority Inheritance Protocol limits the priority inversion phenomenon, the maximum blocking time for high priority tasks can still be significant, due to possible chained blocking conditions. Moreover, deadlock can occur if semaphores are not properly used in nested critical sections.

Priority Ceiling Protocol

The Priority Ceiling Protocol [21] provides a better solution for the priority inversion phenomenon, also avoiding chained

\[
\forall L > 0 \quad \sum_{i=1}^{n} \left[ \frac{L + T_i - D_i}{T_i} \right] C_i \leq L. \quad (3)
\]

\[
L^* = \frac{\sum_{i=1}^{n} (T_i - D_i) C_i}{(1 - U) T_i}. \quad (4)
\]
blocking and deadlock conditions. The basic idea behind this protocol is to ensure that, whenever a task $\tau$ enters a critical section, its priority is the highest among those that can be inherited from all the lower priority tasks that are currently suspended in a critical section. If this condition is not satisfied, $\tau$ is blocked and the task that is blocking $\tau$ inherits $\tau$’s priority. This idea is implemented by assigning each semaphore a priority ceiling equal to the highest priority of the tasks using that semaphore. Then, a task $\tau$ is allowed to enter a critical section only if its priority is strictly greater than all priority ceilings of the semaphores locked by the other tasks. As for the Priority Inheritance Protocol, the inheritance mechanism is transitive. The Priority Ceiling Protocol, besides avoiding chained blocking and deadlocks, has the property that each task can be blocked for at most the duration of a single critical section.

**Schedulability Analysis**

The importance of the protocols for accessing shared resources in a real-time system derives from the fact that they can bound the maximum blocking time experienced by a task. This is essential for analyzing the schedulability of a set of real-time tasks interacting through shared buffers or any other non-preemptable resource, e.g., a communication port or bus. To verify the schedulability of task $\tau_i$ using the processor utilization approach, we need to consider the utilization factor of task $\tau_i$, the interference caused by the higher priority tasks and the blocking time caused by lower priority tasks. If $B_i$ is the maximum blocking time that can be experienced by task $\tau_i$, then the sum of the utilization factors due to these three causes cannot exceed the least upper bound of the scheduling algorithm, that is:

$$\forall i, 1 \leq i \leq n, \sum_{k \in hp(i)} \frac{C_k}{T_k} + \frac{B_i}{T_i} \leq i(2^{\frac{1}{i}} - 1). \quad (5)$$

where $hp(i)$ denotes the set of tasks with priority higher than $\tau_i$. The same test is valid for both the protocols described above, the only difference being the amount of blocking that each task may experience.

**Asynchronous communication buffers**

This section describes how blocking on shared resources can be avoided through the use of Cyclical Asynchronous Buffers [5], or CABs, a kind of wait free mechanism which allows tasks to exchange data without forcing a blocking synchronization.

In a CAB, read and write operations can be performed simultaneously without causing any blocking. Hence, a task can write a new message in a CAB while another task is reading the previous message. Mutual exclusion between reader and writer is avoided by means of memory duplication.

In other words, if a task $\tau_W$ wants to write a new message into a CAB while a task $\tau_R$ is reading the current message, a new buffer is created, so that $\tau_W$ can write its message without interfering with $\tau_R$. As $\tau_W$ finishes writing, its message becomes the most recent one in the CAB.

To avoid blocking, the number of buffers that a CAB must handle has to be equal to the number of tasks that use the CAB plus one. CABs were purposely designed for the cooperation among periodic hard tasks. In general, a CAB provides a one-to-many communication channel, which at any instant contains the latest message or data inserted in it. A message is not consumed by a receiving process, but is maintained into the CAB structure until a new message is overwritten. As a consequence, once the first message has been put in a CAB, a task can never be blocked during a receive operation. Similarly, since a new message overwrites the old one, a sender can never be blocked. Notice that, using such a semantics, a message can be read more than once if the receiver is faster than the sender, while messages can be lost if the sender is faster than the receiver. However, this is not a problem in many control applications, where tasks are interested only in fresh sensory data rather than in the complete message history produced by a sensory acquisition task.

To insert a message in a CAB, a task must first reserve a buffer from the CAB memory space, then copy the message into the buffer, and finally put the buffer into the CAB structure, where it becomes the most recent message. This is done according to the following scheme:

```c
buf_pointer = reserve(cab.id);
<copy message in *buf_pointer>
putmes(buf_pointer, cab.id);
```

Similarly, to get a message from a CAB, a task has to get the pointer to the most recent message, use the data, and then release the pointer. This is done according to the following scheme:

```c
mes_pointer = getmes(cab.id);
<cuse message>
unget(mes_pointer, cab.id);
```

**An example**

To better illustrate the CAB mechanism, we describe an example in which a task $\tau_W$ writes messages in a CAB, and two tasks, $\tau_R1$ and $\tau_R2$, read messages from the CAB. As it will be shown below, to avoid blocking and preserve data consistency, the CAB must contain 4 buffers. Consider the following sequence of events:

- At time $t_1$, task $\tau_W$ writes message $M_j$ in the CAB. When it finishes, $M_j$ becomes the most recent data (mrd) in the CAB.
- At time $t_3$, task $\tau_{R1}$ asks the system to read the most recent data in the CAB and receives a pointer to $M_1$.
- At time $t_4$, task $\tau_{W}$ asks the system to write another message $M_2$ in the CAB, while $\tau_{R1}$ is still reading $M_1$. Hence, a new buffer is reserved to $\tau_{W}$. When it finishes, $M_2$ becomes the most recent data in the CAB.
- At time $t_5$, while $\tau_{R1}$ and $\tau_{R2}$ are still reading, $\tau_{W}$ asks the system to read the most recent data in the CAB and receives a pointer to $M_2$.
- At time $t_6$, while $\tau_{R1}$ and $\tau_{R2}$ are still reading, $\tau_{W}$ asks the system to write a new message $M_3$ in the CAB. Notice that, in this situation, $M_3$ cannot be overwritten (being the most recent data), hence a fourth buffer must be reserved to $\tau_{W}$. In fact, if $M_3$ is overwritten, $\tau_{W}$ could ask reading the CAB while $\tau_{W}$ is writing, thus finding the most recent data in an inconsistent state. When $\tau_{W}$ finishes writing $M_3$ into the fourth buffer, the mrd pointer is updated and the third buffer can be recycled if no task is accessing it.
- At time $t_7$, $\tau_{R1}$ finishes reading $M_3$ and releases the first buffer (which can then be recycled).
- At time $t_8$, $\tau_{R1}$ asks the system to read the most recent data in the CAB and receives a pointer to $M_4$.

Since, in general, it is very difficult to bound a priori the number of interrupts that a task may experience, the delay introduced by the interrupt mechanism on tasks’ execution becomes unpredictable.

To reduce the interference of the drivers on the application tasks and still perform I/O operations with the external world, the operating system should handle interrupts with appropriate aperiodic service mechanisms that can be analyzed to provide a priori guarantee of the real-time constraints [5]. In other words, interrupts should be handled as soon as possible, but without jeopardizing the schedulability of critical control activities.

A typical technique used in real-time systems to schedule aperiodic requests is based on the concept of aperiodic server, which is a kind of periodic task whose purpose is to serve aperiodic requests in a predictable fashion. Like a periodic task, a server is characterized by a period $T_s$ and a budget $C_s$. In general, the server is scheduled with the same algorithm used for periodic tasks, and, once active, it serves the aperiodic requests within the limit of its budget. The order of service of the aperiodic requests is independent of the scheduling algorithm used for the periodic tasks, and it can be a function of the arrival time, computation time or deadline.

During the last years, several aperiodic service algorithms have been proposed in the real-time literature, differing in performance and complexity. Under fixed priority schemes, the most common algorithms are the Polling Server [14], the Deferrable Server [25], and the Sporadic Server [23]. Under dynamic priority schemes (which are more efficient in the average) we recall the Total Bandwidth Server [24] and the Constant Bandwidth Server [1].

**Conclusions**

This paper illustrated the importance of a suitable operating system support when executing complex control applications subject to real-time constraints. Among the most important features needed to execute control applications with high predictability, a real-time kernel should provide:
- a proper support for executing periodic tasks with explicit timing constraints;
- an efficient scheduling algorithm to maximize the processor utilization and enable the user to analyze and verify the schedulability of the task set;
- a runtime mechanism to detect execution overruns and deadline misses, to allow the application to react with proper recovery actions;
- suitable concurrency control protocols to correctly access shared resources;
- non blocking communication mechanisms for exchanging information among periodic tasks running at different rates;
- and appropriate aperiodic service mechanisms to handle interrupt drivers in a controlled fashion and protect the execution of critical tasks.

Without these features, complex control applications could experience highly variable and unpredictable delays, which
could significantly degrade the performance of the controlled system, or even jeopardize its stability. Today, unfortunately, only a few research operating systems provide full real-time features able to cope with the problems presented in this paper. For example, VxWorks [26] and QNX-Neutrino [18], the most popular commercial real-time kernels, include the Priority Inheritance protocol for avoiding priority inversion, but do not have specific support for handling periodic tasks with explicit timing constraints, do not support deadline scheduling, do not have a time accounting mechanism for measuring the actual execution time of tasks (necessary to handle execution overruns and overload conditions), and do not include advanced aperiodic service algorithms. All these features are included into Shark [22], a free modular real-time kernel for PC platforms, developed at the Scuola Superiore Sant’Anna of Pisa. Each function (like scheduling, resource management policy, synchronization protocol, aperiodic service algorithm, etc.) is available as a software module that can be easily replaced, or combined with other modules, to build a custom kernel for specific real-time applications. Other kernels providing some of the real-time features mentioned in this paper are MarteOS [20] and Erika Enterprise [10]. MarteOS is a real-time kernel for PC platforms developed at the University of Cantabria (Spain); it is mostly written in ADA, has a POSIX interface, and allows the user to specify the scheduler at the application level. Erika Enterprise is an OSEK compliant real-time kernel specifically designed for minimal embedded systems with limited onboard resources; it is fully configurable from a minimal memory footprint of 2 Kbytes, up to more complete configurations.

References

[22] The Shark real-time kernel. URL: http://shark.sssup.it/