Multitasking in Embedded System Designs

https://doi.org/10.3991/ijoe.v15i09.10631

Radoslav Mavrevski, Metodi Traykov,
South-West University, Blagoevgrad, Bulgaria
radoslav_sm@abv.bg

Ivan Trenchev
South-West University, Blagoevgrad, Bulgaria
University of Library Studies and Information Technologies Sofia, Bulgaria

Abstract—It is common knowledge in Information Technology (IT) that an embedded system is based on microprocessor and is built to control a function or a range of functions. Although, it is not designed to be programmed by the end user in the same way that a PC is, it is designed to perform one particular task with choices and different options [1-5]. Multitasking is a method by which multiple tasks, also known as processes, share common processing resources, such as CPU. The main aim of this paper is analysis of the design of the embedded systems and a focus on mid-level abstractions for concurrent programs.

Keywords—Embedded System, multitasking, design

1 Introduction

The main “parts” that consist an embedded system are: Processor, Memory, Peripherals, Software and Algorithms.

The main criteria for the processor are whether it can provide the processing power needed to execute the tasks within the system. Sometimes it occurs that the tasks are either underestimated in terms of their size and/or complexity or that the specification is beyond the processor’s capabilities. Usually, these issues are in bigger scale because of the performance measurement used to judge the processor. They may execute completely out of cache memory and thus give an artificially high performance level which the final system cannot meet because its software does not fit in the cache. The software overheads for high level languages, operating systems and interrupts may be higher than expected. These are all issues that can turn a paper design into failed reality [2-4]. While processor performance is important and consists the first gating criterion, there are others such as cost, power consumption, software tools and component availability.

Memory is an important part of any embedded system design, especially nowadays that the requirements and the data are huge. The software influences the resources of a computer system and its memory. Memory can determine how the soft-
硬件是设计、编写和开发的。它主要在嵌入式系统中执行两个功能：1）它为要运行的软件提供存储；和2）它为数据（如程序变量和中间结果、状态信息和任何其他数据）提供存储，这些数据可以在操作过程中创建。

嵌入式系统必须能够通过外围设备与外部世界通信。输入外围设备通常与测量外部环境的传感器相关联，从而有效地控制嵌入式系统执行的操作[2-4]。因此，嵌入式系统可以设计为三个阶段的管道。数据和信息被放入管道的第一阶段；第二阶段执行过程；第三阶段输出结果和数据。

有五种主要类型的外围设备：二进制输出，简单的外部引脚，其逻辑状态可以由处理器控制为零（off）或一（on）；串行输出，是发送或接收数据的接口，使用一种或两种引脚的串行模式；模拟值，是系统和外部环境之间的接口，需要从模拟到数字和反之转换；显示，可以是简单的LED和七段显示到小型LCD面板；时间衍生输出（定时器和计数器）。

软件组件在嵌入式系统中往往包含为系统增加价值的技术，并定义它做什么以及做得多好[1-9]。软件包括不同的组件：自定义和配置；应用程序（模块）；操作系统；错误处理和维护。

算法是软件组件的关键部分，使嵌入式系统的行为方式与之匹配，以便于其功能。它们可以从数学处理到外环境的模型，用于解释和利用来自外部传感器的信息，并生成控制信号。如今，通过数字技术，编码模拟数据的算法由标准机构定义。正确的实施非常重要，例如，它可能允许以更便宜的硬件执行相同的函数（效率）。由于大多数嵌入式系统旨在商业上成功，所以选择“正确”的算法极其重要。

2 Material and Method

在进一步讨论多任务之前，有必要解释操作系统的概念。它是提供缓冲的软件环境，用户和低级接口与硬件之间的接口。它提供了一个常数界面和一组工具，使用户可以快速有效地利用系统。此外，软件允许从一个系统移动到另一个系统，从而使应用程序程序硬件独立。通常情况下包含调试工具。因此，测试过程可以在较短的时间内完成。

Before we move further on, to the part of multitasking, it is necessary to explain what an operation system is. It is a software environment that provides a buffer between the user and the low-level interfaces to the hardware within a system. It provides a constant interface and a set of utilities to enable users to utilize the system quickly and efficiently [7]. Furthermore, software is allowed to be moved from one system to another and thus can make application programs hardware independent. Very often there are program debugging tools included. Therefore, the testing process is completed in lesser time.
Most of the embedded systems nowadays demand a multitasking operating system. It is an operating system that can run multiple applications simultaneously and provide intertask communication and control. A multitasking operating system works by dividing the processor’s time into discrete time slots. Each application or task demands a specific number of time slots to complete its execution. The operating system kernel decides which task can have the next slot, so instead of a task executing continuously until completion, its execution is interleaved with other tasks. This sharing of processor time between tasks gives the illusion to each user that he is the only one using the system [8].

Further on, comes the analysis of the mechanisms that are used in software in order to offer concurrent execution of sequential code. With this, multitasking can be achieved. There are multiple reasons for executing more than one sequential program concurrently and they all involve timing. One of the main reasons is to improve responsiveness by avoiding situations where long-running programs can block a program that responds to external stimuli, such as sensor data or a user request [1-3]. Faster responsiveness reduces latency, which is in fact, the time between the occurrence of a stimulus and the response. Moreover, another reason is to improve performance by allowing a program to be executed simultaneously on multiple cores or processors. Furthermore, one other reason is to directly control the order of external actions. A program may need to perform some action, such as updating a display or saving data that was recently added/modified, at particular times, no matter what other tasks might be executing at the same time.

Figure 1 shows that multitasking are in fact mid-level techniques which are directly related with the low-level and high-level mechanisms. Embedded system designers usually use these mid-level mechanisms to build applications, but it is becoming increasingly common for designers to use the high-level mechanisms instead. The designer builds a model of computation using a software tool. This model is then automatically or semi-automatically translated into a program that exploits the mid-level or low-level mechanisms. The process of translation is widely known as code generation or auto coding.
The mechanisms are provided by an operating system, a microkernel, or a library of procedures. They can be rather difficult to implement correctly, and because of that, the work should be done by very experienced people (experts).

3 Results and Discussion

When it comes to the programming language that application programmers use to develop software, a language that expresses a computation as a sequence of operations is called an imperative language. C is an example of imperative language. In Figure 2 there is an example of a program written in C which implements a commonly used design pattern which is known as the observer pattern [10].

```
1 #include <stdio.h>
2 #include <stdlib.h>
3 int x; // Value that gets updated.
4 typedef void notifyProcedure(int); // Type of notify proc.
5 struct element {
6   notifyProcedure* listener; // Pointer to notify procedure.
7   struct element* next; // Pointer to the next item.
8 };
9 typedef struct element* element_t; // Type of list elements.
10 element_t* head = 0; // Pointer to start of list.
11 element_t* tail = 0; // Pointer to end of list.
12
13 // Procedure to add a listener.
14 void addListener(notifyProcedure* listener) {
15   if (head == 0) {
16     head = malloc(sizeof(element_t));
17     head->listener = listener;
18     head->next = 0;
19     tail = head;
20   } else {
21     tail->next = malloc(sizeof(element_t));
22     tail = tail->next;
23     tail->listener = listener;
24     tail->next = 0;
25   }
26 }
27 // Procedure to update x.
28 void update(int newx) {
29   x = newx;
30   if (Notify listeners.
31     element_t* element = head;
32     while (element != 0) {
33       (*element->listener)(newx);
34       element = element->next;
35     }
36   }
37 // Example of notify procedure.
38 void printf(int arg) {
39   printf("%d", arg);
40 }
```

Fig. 2. An example of a program written in C (observer pattern)

In the following example, an update procedure changes the value of a variable x. The other programs or other parts of the program, which are called “observers”, will be notified every time x is changed by calling a callback procedure. For instance, the value of x might be displayed by an observer on a screen. Whenever the value chang-
es, the “observer” needs to be notified so that it can update the display on the screen. The example below demonstrates a main procedure that uses the procedures defined in Figure 2:

```c
int main(void) {
    addListener(&print);
    update(1);
    addListener(&print);
    update(2);
    return 0;
}
```

In this example, the program registers the `print` procedure as a callback twice, then performs an update, changing the value of x (x = 1), then registers the `print` procedure again, and finally does another update, changing the value of x (x = 2). The `print` procedure prints the current value of the variable x, so the final output when executing this program will be 1 1 2 2 2.

In general, a program in C specifies an order of steps, where each step changes the state of the memory in the machine. In C, the state of the memory is represented by the values of variables [1-4].

In the example which was described in Figure 2, the state of the memory has the value of variable x which is a global variable. A global variable is visible through all the procedures and “parts” of the program. Furthermore, there is a list of elements pointed to by the variable head, which consists another global variable. The list itself is represented as a linked list, where each element in the list contains a function pointer referring to a procedure to be called when x changes [1-4].

In computer science, a linked list is a linear collection of data elements, called nodes, each pointing to the next node by means of a pointer. It is a data structure consisting of a group of nodes which together represents a sequence. Under the simplest form, each node is composed of data and a reference (in other words, a link) to the next node in the sequence. This structure allows for efficient insertion or removal of elements from any position in the sequence during iteration. More complex variants add additional links, allowing efficient insertion or removal from arbitrary element references. Below, there is a definition of a linked list for the program in Figure 2:

```c
typedef void notifyProcedure(int);
struct element {
    notifyProcedure* listener;
    struct element* next;
};
typedef struct element element_t;
element_t* head =0;
element_t* tail =0;
```
These are all examples of code that runs sequentially, line after line. Complex programs can create difficulties handing values and the state of memory. The problems get much worse when concurrency gets into the mix. C programs with mid-level concurrency mechanisms such as threads are extremely difficult and prone to errors.

Threads are imperative programs that run concurrently and share a memory space. One of the advantages is the fact that the programs can access each other’s variables. The term "threads" is widely used to refer to particular ways of constructing programs that share memory. In this paper however, the term will refer to any mechanism where imperative programs run concurrently and share memory.

Most operating systems provide a higher-level mechanism which is provided in the form of a series of procedures that a programmer can exploit. Such procedures typically conform to a standardized API (application program interface), which makes it possible to write programs that are portable, which in fact is very useful as they run on multiple processors and/or multiple operating systems. Pthreads (or POSIX threads) is such an API. It is integrated into many modern operating systems. Pthreads defines a set of C programming language types, functions and constants [1-4]. In Figure 3, there is an example of a multithreaded program.

```c
#include <stdio.h>

void printN(void* arg) {  
int i;
for (i = 0; i < 10; i++) {  
 printf("My ID: %d\n", *(int*)arg);
}

return NULL;
}

int main(void) {  

pthread_t threadID1, threadID2;
void* exitStatus;
int x1 = 1, x2 = 2;

pthread_create(&threadID1, NULL, printN, &x1);
pthread_create(&threadID2, NULL, printN, &x2);

printf("Started threads in\n");

pthread_join(threadID1, &exitStatus);
pthread_join(threadID2, &exitStatus);
return 0;
}
```

**Fig. 3.** Multithreaded program with Pthreads

The `printN` procedure, the procedure that the thread begins executing, is known as the start routine. It prints the argument passed to it 10 times and then exits, which will force the thread to end. The `main` procedure creates two threads; each of one will execute the start routine. The first one, which is created on line 14, will print the value 1. The second one, which is created on line 15, will print the value 2. During the execution of the program, values 1 and 2 will be the outputs that will be displayed on screen, in some mixed order that depends on scheduler of the thread. Each time the program runs, the results will be different orders of 1's and 2's.

In addition, the `pthread_create` procedure creates a thread and returns immediately. The start routine may or may not have actually started running when it returns. Lines
17 and 18 use pthread join to ensure that the main program does not terminate before the threads have finished. Without these two lines, running the program may not yield any output at all from the threads [1-5]. Generally, a start routine may return or not. In embedded applications, it is common to define start routines that never return. For instance, the start routine might run forever and produce an output display periodically. If the start routine does not return, then any other thread that calls its pthread join will be blocked indefinitely.

The center part of an implementation of threads is a scheduler that chooses which thread will be the one executed next when a processor is available to execute one of them. The choice may be based on fairness, where the principle is to give every active thread an equal opportunity and time to run, on timing limits, or in terms of how important it is according to its priority. The first matter that concerns the programmers is how and when the scheduler is invoked. A simple technique known as cooperative multitasking does not interrupt a thread unless the thread itself calls a specific procedure or one out of a set of procedures. For instance, the scheduler may intervene each time an operating system service is invoked by the currently executing thread. Respectively, an operating system service is invoked by calling a library procedure. Each thread has its own stack, and when the procedure call is made, the return address will be pushed onto the stack. If the scheduler determines that the currently executing thread should continue to execute, then the requested service is completed and the procedure returns as normal. On the other hand, if the scheduler determines that the thread should be interrupted and another thread should be selected for execution, then instead of returning, the scheduler makes a record of the stack pointer of the currently executing thread, and then changes the stack pointer to point to the stack of the selected thread. Moreover, it returns as normal by popping the return address off the stack and resuming execution, but this time in a new thread.

Concerning the main drawback of cooperative multitasking, it is that a program may run for a long time without making any operating system service calls, in which case other threads will be put to hold almost indefinitely. To avoid this, most of the operating systems out there, offer an interrupt service routine that runs at certain time intervals. This routine maintains a system clock, which provides application programmers with a solution to obtain the current time of day and enables periodic invocation of the scheduler with the use a timer interrupt.

In multitasking, two concurrent pieces of code race to access the same resource, and the exact order in which their accesses occur affects the results of the program. Not all race conditions are as bad as the previous example, where some outcomes of the race cause catastrophic failure. One effective solution to avoid these disasters is by taking advantage of a mutual exclusion lock (or mutex). The following piece of code is such an example:

```c
pthread_mutex_t lock = PTHREAD_MUTEX_INITIALIZER;

void addListener(notifyProcedure* listener) {
    pthread_mutex_lock(&lock);
    if (head == 0) {
```

http://www.i-joe.org
On the first line a global variable called lock is created and initialized. Inside the addListener procedure acquires the lock. The principle is that only one thread can hold the lock each time. The pthread_mutex_lock procedure will block until the calling thread can acquire the lock. When the procedure addListener is called by a thread and begins executing, pthread mutex lock does not return until no other thread holds the lock. Once it returns, this calling thread holds the lock. The pthread_mutex_unlock call at the end releases the lock. It is a serious error in multithreaded programming to fail to release a lock.

A mutual exclusion lock prevents two threads from simultaneously accessing or changing a shared resource of the memory. The code between the lock and unlock is a critical section. At any cost, only one thread can be executing code in such a critical section. A programmer may need to ensure that all accesses to a shared resource are similarly protected by locks [1-4].

However, the mutex added in the previous example is not enough to avoid this problem. The mutex does not prevent thread A from being halted. So, there is need to protect all accesses of the data structure with mutexes, which can be done with altering update procedure as described below:

```c
void update(int  newx) {
    x = newx;
    // Notify listeners.
    pthread_mutex_lock(&lock);
    element_t* element = head;
    while (element != 0) {
        (*(element->listener))(newx);
        element = element->next;
    }
    pthread_mutex_unlock(&lock);
}
```

The above code will refrain the update procedure from reading the list data structure while it is being altered by any other thread.

As mutex can be used frequently in programs, the risk of deadlock gets bigger. A deadlock occurs when some threads become permanently halted struggling to acquire locks. As an example, if thread A holds lock1 and then stops trying to acquire lock2, which is taken by thread B, and then thread B stops trying to get lock1. This will lead the system to crash and the program needs to be aborted.

Deadlock can be rather difficult to avoid. One simple technique is to use only one lock throughout an entire multithreaded program. On the other hand, that technique
does not launch very modular programming. In addition, it can make it difficult to meet real-time constraints because some shared resources (e.g., displays) may need to be held long enough to cause deadlines to be missed in other threads [1-3].

Sometimes in a very simple microkernel, it is easy to use the enabling and disabling of interrupts as a single global mutex. If there is a single processor and that interrupts are the only available mechanism by which a thread may be halted, then disabling interrupts avoids the suspension of a thread. On the other hand, in most Operating Systems, threads can be halted for various reasons, thus this technique won't give back.

Moving further, another technique is to ensure that when there are more than one mutex locks, each thread acquires the locks in the same order. This can be difficult to guarantee, however, for several reasons:

- Most programs are written by multiple people, and the locks acquired within a procedure are not part of the signature of the procedure. Thus, this technique relies on very careful and constant cooperation and writing down throughout a programming team. Everytime a lock is added, and then all parts of the program that acquire locks may have to be changed, to sustain its “vitality”.

- Correction of the source code may be proved to be extremely difficult. If a programmer wants to call a procedure that acquires lock1, then it must first release any locks it holds. As soon as it releases those locks, it may be halted, and the resource that it held those locks to protect may be modified. Once it has acquired lock1, it must then reacquire those locks, but it will then need to assume it no longer knows anything about the state of the resources, and it may have to redo considerable work.

Of course, there are many other ways to avoid deadlock. For example, a particularly elegant technique synthesizes constraints on a scheduler to prevent deadlock [9]. However, most of them either inflict serious constraints and problems on the programmer or demand considerable sophistication to apply, which suggests that the problem may be with the concurrent programming model of threads [1-5].

Threads also suffer from problems that have to do with the memory model of the programs. Implementation of threads gives some kind of memory consistency model, which defines how variables that are read and written by various threads appear to them. Reading a variable should return the last value written to the variable. For instance, in a program, all variables are initialized with value 0 (zero), and thread A executes the following two lines:

1 x = 5;
2 w = y;

while thread B executes the following two lines:

1 y = 5;
2 z = x;

Then, after both threads have executed these lines, the expected result is that at least one of the two variables w and z will have value 5 (five). This is known as sequential consistency [10]. Sequential consistency means that the result of any execution is the same as if the operations of all threads are executed in some sequential
order, and the operations of each individual thread appear in this sequence in the order
specified by the thread.

On the other hand, sequential consistency is not working everytime by most
implementations of Pthreads. In fact, such a guarantee is rather difficult to be achieved
on modern processors and modern compilers. A compiler, for instance, is free to re-
order the instructions in each of these threads because there is no dependency between
them. There is also a chance that the hardware might reorder them instead. A solution
to this is to very carefully guard such accesses to shared variables using mutual exclu-
sion locks.

Multithreaded programs can be very difficult to understand. Furthermore, it can be
difficult to trust the programs because problems in the code may not show up in the
testing process. A program may have the possibility of deadlock, for example, but
nonetheless run correctly for years without the deadlock ever appearing [1-5].

```c
void update(int newx) {
    x = newx;
    // Copy the list
    pthread_mutex_lock(&lock);
    element_t *head = NULL;
    element_t *tail = NULL;
    element_t *element = head;
    while (element != 0) {
        if (head == NULL) {
            head = malloc(sizeof(element_t));
            head->listener = head->listener;
            head->next = 0;
            tail = head;
        } else {
            tail->next = malloc(sizeof(element_t));
            tail = tail->next;
            tail->listener = element->listener;
            tail->next = 0;
        }
        element = element->next;
    }
    pthread_mutex_unlock(&lock);
    // Notify listeners using the copy
    element = head;
    while (element != 0) {
        (*element->listener)(newx);
        element = element->next;
    }
}
```

**Fig. 4.** Modified update procedure

In Figure 4, there is a modified version of the *update* procedure that was described
earlier. This code does not hold lock when it calls the listener procedure. Instead, it
holds the lock while it constructs a copy of the list of the listeners, and then it releases
the lock. After releasing the lock, it uses the copy of the list of listeners to notify the
listeners. However, it carries a potentially critical problem that is hard, if not com-
pletely impossible, to be detected in testing. For example, thread A calls *update* with
newx = 0, indicating "All systems good.". A gets suspended just after releasing the
lock, but before performing the notifications. While it is suspended, thread B calls
$update$ with $newx = 1$, indicating "Emergency! The engine is off!" Suppose that this
call to update completes before thread A gets a chance to resume. When thread A
resumes, it will notify all the listeners, but it will notify them of the wrong value! If
one of the listeners is updating a pilot display for an aircraft, the display will indicate
that all systems are normal, when in fact the engine is on fire [1-3]. This is an exam-
ple of a very serious issue that could be fatal (loss of human lives) and thus it must be
solved effectively.

Last but not least, processes are also important to keep an eye on when it comes to
embedded systems. Processes are imperative programs with their own memory spac-
es. One of their main characteristics is that they cannot refer to each other’s variables,
and consequently they do not demonstrate the same difficulties as threads. Communi-
cation between the programs must occur through the use of mechanisms provided by
the operating system, a library, or microkernel.

When it comes to the implementation, processes generally demand hardware sup-
port in the form of a memory management unit or MMD. The MMD protects the
memory of one process from accidental, not desired reads or writes by another pro-
cess and also provides address translation, giving each process the “illusion” of a
fixed memory address space that is the same for all processes. When a process ac-
cesses a memory location in that address space, the MMD shifts the address to refer to
a location in the portion of physical memory allocated to that process [1-5].

Operating systems offer various mechanisms, often even including the ability to
create shared memory spaces, which of course opens the programmer to all the poten-
tial difficulties of multithreaded programming. After all, in order to achieve concurren-
cy, the processes must able to communicate. A flexible mechanism for communi-
cating between processes is message passing. One process creates a chunk of data,
deposits it in a carefully controlled section of memory that is shared, and then notifies
other processes that the message is ready. Those other processes can block waiting for
the data to become ready. Message passing requires some memory to be shared, but it
is implemented in libraries that are presumably written by experts. An application
programmer invokes a library procedure to send a message or to receive a message.
Below, in Figure 5, there is an example of a simple message passing application.

This program uses a producer-consumer pattern, where one thread produces a se-
ries of messages (a stream), and another thread consumes the messages. This pattern
can be used to implement the observer pattern without deadlock risk and without the
insidious error discussed in the previous section. The update procedure would always
execute in a different thread from the observers, and would produce messages that are
consumed by the observers. The code executed by the producing thread is given by
the producer procedure and the code for the consuming thread by the consumer pro-
cedure. The producer invokes a procedure called send (to be defined) on line 4 to send
an integer-valued message. The consumer uses get (also to be defined) on line 10 to
receive the message. The consumer is assured that get does not return until it has
actually received the message. In this case, consumer never returns, so this program
will not terminate on its own.
Conclusions

To sum up, there has been an analysis of the design of the embedded systems and a focus on mid-level abstractions for concurrent programs. Examples of codes that handle multitasking have been demonstrated to provide a look on threads, which are sequential programs that execute concurrently and share variables and memory. However, threads are tricky to handle and require attention and study because there are several issues, such as mutual exclusion and deadlock. Message passing schemes avoid some of the difficulties, but not all of them, at the expense of being somewhat more constraining by prohibiting direct sharing of data. In the future, designers and programmers need to focus on using higher levels of abstraction (see Figure 1).

Acknowledgement

This research would not have been possible without the financial assistance of the following project: "Application of the mixed reality in the training and promotion of the cultural heritage for the purposes of the in the university information environment" financed by National Science Fund of the Ministry of Education and Science of the republic of Bulgaria with Contract № KP – 06 – OPR 05/14 from 17.12.2018, led by Prof. DSc Irena Peteva.

```c
void* producer(void* arg) {
    int i;
    for (i = 0; i < 10; i++) {
        send();
    }
    return NULL;
}

void* consumer(void* arg) {
    while(1) {
        printf("received %d", get());
    }
    return NULL;
}

int main(void) {
    pthread_t threadID1, threadID2;
    void* exitStatus;
    pthread_create(&threadID1, NULL, producer, NULL);
    pthread_create(&threadID2, NULL, consumer, NULL);
    pthread_join(threadID1, &exitStatus);
    pthread_join(threadID2, &exitStatus);
    return 0;
}
```

Fig. 5. Example of a simple message-passing application
6 References


7 Authors


Metodi Traykov is Assistant in Department of Informatics, Faculty of Mathematics and Natural Sciences, member of University Center for Advanced, Bioinformatics Research, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., Blagoevgrad, Bulgaria. PhD on Informatics. Scientific Interest: programming and bioinformatics. He is one of the organizers of the XXVII REPUBLICAN STUDENT PROGRAMMING OLYMPIAD, 2015, http://bcpc.eu/XXVII/.
Ivan Trenchev is Associate professor in Department of Electrical Engineering, Electronics and Automation, Faculty of Engineering, member of University Center for Advanced, Bioinformatics Research, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., Blagoevgrad, Bulgaria and Associate professor in University of Library Studies and Information Technologies Sofia, Bulgaria. PhD on Informatics. Scientific interest: virtual reality (VR), computer modelling and bioinformatics. He is one of the organizers of the XXVII REPUBLICAN STUDENT PROGRAMMING OLYMPIAD, 2015, http://bcpc.eu/XXVII/.