Seventh College on Microprocessor-Based Real-Time Systems in Physics

Abdus Salam ICTP, Trieste.

October 28 – November 22, 2002

Toward Real-time Linux

by Catharinus Verkerk

Contents

Chapter 1

Toward Real-time Linux

by Catharinus Verkerk

Abstract

We examine the operating system support needed for a real-time application. We'll see to what extent Linux satisfies the requirements and what has been done to adapt it.

1.1 Introduction and a few definitions

A *real-time system* is defined as a system that *responds to an external stimulus within a specified, short time*. This definition covers a very large range of systems. For instance, a data-base management system can justly claim to operate in real-time, if the operator receives replies to his queries within a few seconds. As soon as the operator would have to wait for a reply for more than, say, 5 seconds, she would get annoyed by the slow response and maybe she would object to the adjective "real-time" being used for the system. Apart from having unhappy users, such a slow data-base query system would still be considered a real-time system.

The real-time systems we want to deal with are much more strict in requiring short response times than a human operator is, generally speaking. Response times well below a second are usually asked for, and often a delay of a few milliseconds is already unacceptable. In very critical applications the response may even have to arrive in a few tens of microseconds.

In order to claim rightly that we are having a real-time system, we must specify the response time of the system. If this response time can be occasionally exceeded, without any real harm being done, we are dealing with a *soft real-time system*. On the contrary, if it is considered to be a *failure* when the system does not respond within the specified time, we are having a *hard real-time system*. In a hard real-time system, exceeding the specified response time may well result in serious damage of one sort or another, or in extreme cases even in the loss of human life. A data-base query system will generally fall in the first category: it will make little difference if a human operator will have to wait occasionally 6

seconds, instead of the specified 5 seconds response time, and nobody will dare to speak of a failure, as long as the replies to the queries are correct. This does not mean that all data-base systems are soft real-time systems: a data-base may well be used inside a hard real-time system, and its response may become part of the overall reaction time of the system.

Data-base systems are not at the centre of our attention in this course; we rather are interested in systems which control the behaviour of some apparatus, machinery, or even an entire factory. We call these *real-time control systems*. We are literally surrounded by such real-time control systems: video recorders, video cameras, CD players, microwave ovens, and washing machines are a few domestic examples. In the more technical sphere we will find the control of machine tools, of various functions of a car, of a chemical plant, etc., but also automatic pilots, robots, driver-less metro-trains, control of traffic-lights, and many, many more. Several of those systems are hard real-time systems: the automatic pilot is a good example.

We implicitly assumed that the systems we are dealing with are computer controlled. We are in fact interested in investigating the role the computer plays, what constraints are imposed by the part of the system external to the computer, or the environment in general, and what these constraints imply for the program that steers the entire process. We will pay particular attention to the role the underlying operating system plays and to what extent it may help in the development and or running of a real-time control system.

At this point we should define two classes of real-time systems. On the one hand we have *embedded systems*, where the controlling microprocessor is an integral part of the entire product, invisible to the user and where the complete behaviour of the system is factory defined. The user can only issue a very limited and predefined set of instructions, usually with the help of switches, push-buttons and dials. There is no alpha-numeric keyboard available to give orders to the device, nor is there a general output device which can give information on the state of the system. On a washing machine we can select four or five different programs, which define if we will wash first with cold and then with warm water, or if we skip the first, or which define how often we will rinse, if we will use the centrifugal drying or not, etc. If we add the control the user has over the temperature of the water, we have practically exhausted the possibilities of user intervention. The microprocessor included in the system has been programmed in the factory and cannot be reprogrammed by the user. Cost has been the overriding design consideration, user convenience played a secondary or tertiary role. These embedded systems run a monolithic, factory defined program and there is no trace of an interface to an operating system which would allow a user to intervene. This does not mean that such an embedded system does not take account of a number of principles, which should not be neglected in a system that claims to operate in real-time. All real-time aspects are folded into the monolithic program, indistinguishable of the other functions of the program.

The other class of real-time control systems comprises those systems that make use of a normal computer, which has not been severed of its keyboard and of its display device and where a human being can follow in some detail

how the controlled process is behaving and where he can intervene by setting or modifying parameters, or by requesting more detailed information, etc. The essential difference with an embedded system is that a system in this second class can be entirely reprogrammed, if desired. Also, in contrast to an embedded system, the computer is not necessarily dedicated to the controlled process, and its spare capacity may be used for other purposes. So, a secretary may type and print a letter, while the computer continues to control the assembly line.

It is obvious that the latter class of real-time control systems needs to run an operating system on the control computer. This operating system must be aware that it is controlling external equipment and that several operations initiated by it may be time-critical. The operating system must therefore be a **real-time operating system**. We will see in these lectures what this implies for the design and the capabilities of the operating system. We should keep in mind that we speak of generic real-time systems and generic real-time operating systems. The real-time control system does not necessarily use all features of the operating system, but the unused ones remain present, ready to be used at a possible later upgrade of the control system. This again is in contrast with the embedded system, where the parts of the operating system needed are cast in concrete inside the controlling program and where all other parts of it have been discarded.

1.2 The ingredients of a real-time computer controlled system

In order to investigate to some extent what the ingredients of a real-time control system are and what the implications are for a supporting operating system, we will take a simple example, which does not require any a-priori knowledge: a railway signalling system.

Safety in a railway system, and in particular collision-avoidance is based on a very simple principle. A railway track, for instance connecting two cities, is divided into sections of a few kilometers length each (the exact length depends on the amount of traffic and the average speed of the trains). Access to a section — called a block in railway jargon — is protected by a signal or a semaphore: when the signal exhibits a red light, access to the block is prohibited and a train should stop. A green light indicates that the road is free and that a train may proceed. The colour of the light is pre-announced some distance ahead, so that a train may slow down and stop in time. Access to a block is allowed if and only if there is no train already present in the block and prohibited as long as the block is "occupied". Normally all signals exhibit a red light; a signal is put to green only a short time before the expected passage of a train and if the condition mentioned above is satisfied. Immediately after the passage of the train, the signal is put back to red. The previous block is considered to be free only when the entire train has left it.

We will try to outline briefly – and rather superficially – what would be required if we decided to make a centralized, computer controlled system for the signalling of the entire railway system in a small or medium-sized country, comprising a few

thousand kilometers of track, with hundred or so trains running simultaneously. This would be a large-scale system, but it would be conceptually rather simple. The basic rule is: if there is a train moving forward in block $i-1$, and block i is free, the signal protecting the entrance to block i shall be put to green and back to red again as soon as the first part of the train has entered block i. For the time being we consider only double track inter-city connections, where trains are always running in the same direction on a given track.

From the rule we see that we need to know at any instant in time which blocks are free and which are occupied. So we need a sort of a *data-base* to contain this information. This data base must be regularly updated, to reflect faithfully the real situation. In fact, whenever a train is leaving a block and entering another, the data-base must be updated.

How do we know that a train moves from one block to the next? Trains are supposed to run according to a time table and at predefined speeds, so a simple algorithm should be able to provide the positions of all trains in the system at any moment. Unfortunately this assumption is not valid under all circumstances and we need a reliable signalling system, exactly to be able to cope with more or less unexpected situations where trains run too late, or not at all, or where an extra train has been added, or another ran into trouble somewhere. We conclude that it is better to actually *measure* the event that a train crosses the boundary between two blocks. We could put a switch on the rails, which would be closed by the train when it is on top of the switch. We could scan all the switches in our system at regular intervals. How long — or rather how short — should this interval be? A TGV of 200 meter length and running at close to 300 km/h, would be on top of a switch for $2\frac{1}{2}$ seconds. A lonely locomotive, running at 100 km/h would remain on top of the contact for much less than a second. So we must scan some thousand or more contacts in, say $\frac{1}{2}$ second. This can be done, but it would impose a heavy load on the system and we would find the vast majority of the switches open in any case. We could refine our method and scan only those contacts where we expect a train to arrive soon. This would reduce the load on the system, as only hundred or so contacts have to be scanned, but it still is not very satisfactory, as we will continue to find many open contacts. Note that instead of contacts, we could have used other detection methods: strain gauges on the rails, or photo-cells.

A better way of detecting the passage of a train, is by using hardware interrupts [1]. We could generate an interrupt when the contact closes and another when it opens again, indicating the entrance of a train into block i and the exit of the same train from block $i-1$, respectively. We don't lose time then anymore for looking at open contacts. We also simplify the procedure, for we do not have to look anymore at the data-base before the start of a scan, to find out which contacts are likely to be closed by a train soon.

We have discovered here a very important ingredient of any real-time control

[1] For those who may have forgotten: a hardware interrupt is caused by an external electrical signal. The normal flow of the program is interrupted and a jump to a fixed address occurs, where some work is done to *handle* the interrupt. A "return from interrupt" instruction brings us back to the point where the program was interrupted.

system: the *instrumentation with sensors and actuators*. In our case we must sense the presence of a train at given positions along the tracks, and we must actuate the signals, putting them to green and to red again. Generally speaking, the instrumentation of a real-time control system is a very important aspect, which must be carefully considered. Usually, apart from *sensors which provide single-bit information*, such as switches, push-buttons, photocells, *which can also be used to generate hardware interrupts*, we will need *measuring devices, giving an analog voltage output*, which then has to be *converted into a digital value* with an analog-to-digital converter. Conversely, *output devices may be single bit*, such as relays, lamps and the like, or *digital values, to be converted into analog voltages*. **Accuracy, reproducibility, voltage range, frequency response** etc. have to be considered carefully. The operation of a system may critically depend on how it has been instrumented. The *interface to the computer* is another aspect to take into account for its possible consequences. **Speed, reliability and cost** are some of the concurring aspects. We will not dwell any further on these topics in these lectures, as they are too closely related to the particular application, making a general treatment impossible.

For our railway signalling system we mentioned the timetable, claiming that we could not rely on it. We can however use it to check the true situation against it in order to detect any anomaly. These anomalies could then be reported immediately. For instance, we could tell the station master of the destination, that the train is likely to have a delay of x minutes. Another useful thing is to keep a *log* of the situation. This can be used for daily reports to the direction (where they would probably be filed away immediately), but they could prove valuable for extracting statistics and for global improvement of the system. *Operator intervention* is also needed. For instance, when a train, running from station A to station B, leaves station A, it does not yet exist in the data-base of running trains. Likewise, when it arrives at B, it has to be removed from this data-base. This could be done automatically, in principle, but what do we do if it has been decided to run two extra trains, because there is an important football match? We conclude that *data-logging, operator intervention* and some *calculations* (to check actual situation against predicted one) are also essential ingredients of a real-time control system, in addition to the *interrupt handling, interfacing to the sensors and actuators* and *updating of the data-base* reflecting the state of the system.

This idyllic picture of our railway signalling system might stimulate us to start coding immediately. A program which uses the principles outlined above does not seem too difficult to produce. We simply let the program execute a large loop, where all different tasks are done one after the other. The interrupts have made it possible to get rid of a serious constraint, so all seems to be nice and straightforward. Once we would have a first version of the program ready, we would like to test it. Hopefully we will use some sort of a test rig at this stage, and abstain from experimenting with real trains. During the testing stage, we will then quickly wake up and find that we have to face reality.

In our model, we assumed double track connections between cities, where on a given track, trains always run in the same direction. But, even in the case that the entire railway network is double track between cities, we must nevertheless

consider also single track operation, because a double track connection may have to be operated for a limited period of time and for a limited distance as a single track, repair or maintenance work making the other track unusable.

Assume that, on a single track, we have two trains, one in block $k+1$, the other in block $k-1$, running in opposite directions, both toward block k. If we would apply our simple rule, they would both be allowed to enter block k (supposing it was free) and a head-on collision would result. The problem can be solved by slightly modifying our rule: If a train is moving forward in block $i-1$ toward block i, then access to block i will be allowed if blocks i and $i+1$ are free. So both trains will be denied access to block k in our example. We have eliminated the possibility of a head-on collision, but we now have another problem. Assume that our two trains are in block $k-2$ and $k+1$ respectively and running toward each other. Applying our new rule, they would be allowed to enter block $k-1$ and k respectively and both trains would stop, nose to nose at the boundary between these two blocks. We have created a sort of a *deadlock situation*.

The true solution is of course not to allow a south-bound train into an entire section of single track, as long as there is still a north-bound train somewhere in this entire section, and vice-versa. A section consists of several blocks and inside a section there are no switches enabling a train to move from one track to another, nor to put it on a side-track. South-bound and north-bound trains compete for the same "**resource**", the piece of single track railway. They are **mutually exclusive** and only one type (north-bound or south-bound) of train should be allowed to use the resource. If the stretch of single track is long enough, and comprises several blocks, more than one north-bound train can be running on that stretch of track. Now assume that several north-bound trains are occupying the stretch of single track and that a south-bound train presents itself at the northern end of the stretch. It obviously has to wait, but while it is waiting, do we continue to allow more north-bound trains into the stretch? This is a matter of *priority*, which should be defined for each train. A *scheduler* should take the priorities into account and deny the entrance into the stretch for a north-bound train if the waiting south-bound one has higher priority. As soon as the stretch has then been emptied of all north-going trains, the south-bound one can proceed, possibly followed by others.

A similar situation, where two trains may be competing for the same resource, arises when two tracks, coming from cities A and B, merge into a single track entering city C. Obviously, if two trains approach the junction simultaneously, only one can be allowed to proceed, which should be the one with the highest priority. It should be noted that the priority assigned to a train is not necessarily static. It may change dynamically. For instance, a train running behind schedule, may have its priority increased at the approach of the junction and allowed to enter city C, before another train which normally would have had precedence. This latter example illustrates a *synchronization problem*: some trains may carry passengers which have to change trains in city C; the two trains should reach the station of C in the right order.

We have thus discovered some more ingredients (or concepts) for a real-time control system: **priorities, mutual exclusion, synchronization**.

We started off by considering our railway signalling problem being controlled by a single program, which guides all trains through all tracks, junctions and crossings. We have gradually come to have a different look at the problem: *a set of trains, using resources* (pieces of railway track), *and sometimes competing for the same resource*. We can consider our trains as *independent objects*, more or less unaware of the existence of similar objects and of the competition this may imply. In order to get a resource, every train must put forward a request to some sort of a master mind (the real-time operating system), who will honour the request, or put the train in a waiting state.

At this stage, we realize that we better abandon our first version of the program, because it would have to be rewritten from scratch in any case. We have become aware that our particular real-time control system may have many things in common with other real-time systems and that it would be advantageous to take profit from the facilities a real-time operating system offers to solve the problems of mutual exclusion, priorities, etc. Once we have mastered the use of these facilities, we can build on our experience for the implementation of another real-time control system. In case we would obstinately continue to adapt our original program, we would probably find, after months of effort, that we have rewritten large parts of a real-time operating system, but which have been so intimately interwoven with the application program, that it will be difficult, if not impossible, to re-use it for the next application we may be called to tackle.

Other aspects we have not yet considered may also build very nicely on the foundations laid by a real-time operating system. For instance, we have the problem of dealing with *emergencies*. A train may have derailed and obstructed both tracks. Such an unusual and potentially dangerous situation must be immediately notified to the operating system which can then take the necessary measures. If they cannot be notified, a mechanism for detecting potentially dangerous situations must be devised: in our particular case, the system should be alerted if a train does not leave its block within a reasonable time. In other words, a *time-out* could be detected.

Now that we mentioned time, we are reminded of the fact that **time** may play an important role in any real-time system, either in the form of *elapsed time*, or of the *time of the day*. It is difficult to think of a system that could operate without the help of a clock. A **real-time clock** and the possibility to program it to generate a clock interrupt at certain points in time, or after a given time-interval has elapsed, are therefore indispensable ingredients of a real-time control system.

Reliability of the entire system is another item for serious consideration. You certainly do not want a parity error in a disk record to bring your system to a halt or to create a chain of very nasty incidents.

In many cases, we are not dealing with a closed system, so there must be a means of *communicating with other systems* (our national railway network is connected to other networks, and trains do regularly cross the border). *User-friendly interfaces to human operators*, which usually implies the use of graphics, are also very likely to be an essential ingredient of our real-time control system. A large synoptic panel, showing where all trains are in the network, would be the supervisor's dream, not to speak of makers of science fiction films.

In the following lectures, we will investigate in more depth the various features a real-time operating system should provide. Making use of these features will prevent us from re-inventing the wheel.

The question then arises: which real-time operating system should I use? There are several on the market: *OS-9000* for *Motorola 68000 machines*, and *QNX* or *LynxOS* for *Intel machines*, *Solaris* for *Sparc processors*, to mention only a few of the older ones. These systems are sold together with the tools necessary to build a real-time application: **compiler, assembler, shell, editor, simulator, etc**. A minimum configuration would cost US$ 2000-2500, a full configuration may push the price up into the 10 K$ range. This would cause no problem whatsoever for a railway company, but what about you?

Another solution is to use a **real-time kernel**, useful for embedded systems, which you *compile and link into your application*. *VxWorks, MCX11* and *μCOS* are examples. They are much cheaper —or practically free: *MCX11* and *μCOS* [1]—, but you will need a complete development system in addition. This development system could of course be Linux.

The ideal would be to be able to **use Linux for development of a real-time control system, as well as for running the application**. We will see shortly to what extent this is possible at present. Before proceeding, however, we will make sure that we understand the fundamental concept of a **process**.

1.3 Processes

In our example we have seen that a real-time system has a number of tasks to accomplish: besides ensuring that trains could proceed from block to block without making collisions, we had to log data, keep the data-base up-to-date, communicate with the operator, cater for emergency situations, etc. Not all of these tasks have the same priority, of course.

When we analyze a real-time system, we will almost invariably be able to identify *different tasks*, which are more or less *independent of each other*. "Independent of each other" really means that each task can be programmed without thinking too much of the other tasks the system is to perform. At most there is some *intertask communication*, but every task does its job on its own, without requiring assistance from other tasks. If assistance is required, the operating system should provide it. The system designer should identify and define the different tasks in such a way that they really are as independent of each other as possible. Some *synchronization* may be needed: certain tasks can only run after another task has completed. For instance, if some calculations have to be done on collected data, the data collection tasks could be totally separated from the calculation task. In order to make sense, the latter should only be executed when the data collection task has obtained all data necessary for the calculation. This implies that some inter-task communication is needed here. The true difficulty of dividing the overall system requirement up into different tasks consists of choos-

[1] *RTEMS* is a more recent, very complete real-time executive, also available for free.

ing the tasks for maximum independence, or —in other words— for **minimum need of inter-task communication and synchronization**.

These various tasks can now be implemented as different programs and then run as different **processes**.

What exactly is a process and what is the *difference between a program and a process*? A program is an orderly sequence of machine instructions, which could have been obtained by compilation of a sequence of high-level programming language statements. It is not much more than the listing of these statements, which can be stored on disk, or archived in a filing cabinet. It becomes useful only when it is run on a machine and executing its instructions in the desired sequence, thus obtaining some result. It is only useful when it has become a *running process*.

A process is therefore a running (or runnable) program, together with its data, stack, files, etc. It is only when the code of a program has been loaded into memory, and data and stack space allocated to it, that it becomes a runnable process. The operating system will then have set up an entry in the *process descriptor table*, which is also part of the process, in the sense that this information would disappear when the process itself ceases to exist. The operating system may decide at a certain moment to run this runnable process, on the basis of its priority and the priorities of other runnable processes. This would happen in general when the process that is using the CPU is unable to proceed — e.g. because it is waiting for input to become available — or because the time allocated to it has run out.

We should emphasize that we are considering only the case of a single processor system, where only one process can run at a time. The other runnable processes will wait for the CPU to become free again. If the different processes are run in quick succession, a human observer would have the impression that these processes are executed simultaneously.

The consequence of this is that we can write a program to calculate Bessel functions, without having to think at all about the fact that when we will run our program, there may already be fifty or more other processes running, some of them even calculating Bessel functions. In as far as we have written our program to be autonomous, it will not be aware of the existence of other runnable processes in the computer system. Consequently, it cannot communicate with the other processes either: its fate is entirely in the hands of the operating system[1].

There may exist on the disk a general program to calculate Bessel functions and on a general purpose time-sharing computer system several users may be running this program. A reasonable operating system should then keep only one copy of the program code in memory, but each user process running this program should have its *own* process descriptor, its *own* data area in memory, its *own* stack and its *own* files. All users of the computer system will presumably run a **shell**. Command shells, such as *bash* or *tcsh* are very large programs and it

[1] And luckily so: the operating system will also provide protection, avoiding that other processes interfere with ours.

would be an enormous waste if every single user of a time-sharing system would have his own copy of the shell in memory.

In general, we will have a number of runnable processes in our uniprocessor machine, and one process running at a given instant of time. When will the waiting processes get a chance to run? There are two reasons for suspending the execution of the running process: either the *time-slice* allocated to it *has been exhausted*, or *it cannot proceed any further before some event happens*. For instance, the process must wait for input data to become available, or for a *signal* from another process or the operating system, or it has to complete an output operation first, etc. The programmer does not have to bother about this. At a given point in the program, where it needs to have more input data, the programmer simply writes a statement such as: *read(file,buffer,n);*. The compiler will translate this into a call to a *library function*, which in turn will make a **system call**, (or **service request**), which will transfer control to the kernel. Our process becomes *suspended* for the time the kernel needs to process this system call. In the case of a read operation on a file, the kernel will set this into motion, by emitting the necessary orders to the disk controller. As the disk controller will need time to execute this order, the kernel will decide to **block** the execution of the process which was running and which made the system call. This blocked process will be put in the *queue of waiting processes*, and it will become runnable again later, when the disk controller will have notified the kernel —by sending a hardware interrupt— that the I/O operation has been completed. The kernel makes use of the **scheduler** to find, from the queue of runnable processes the one that should now be run. The kernel will then make a **context switch** and this will start our suspended process running.

A context switch is a relatively heavy affair: first all hardware registers of the old (running) process must be saved in the process descriptor of the old process. Then the new process must be selected by the scheduler. If the code and data and stack of the new process are not yet available in memory, they must be loaded. In order to be loaded, it may be necessary first to make room in memory, by swapping out some memory pages which are no longer needed or which are rarely used. The page tables must be updated, and the process descriptors must be modified to reflect the new situation. Finally the hardware registers of our machine must be restored from the values saved at an earlier occasion for the process now ready to start running. The last register to be restored is the program counter. The next machine instruction executed will then be exactly the one where the new process left off when it was suspended the last time.

The execution of a program will thus proceed piecemeal, but without the programmer having to bother about it: the operating system takes care of everything. So the application programmer can continue to believe that his program is the only one in the world. The price to be paid for this convenience is the overhead in time and memory resources introduced by the intervention of the operating system.

For our Bessel function program we are entirely justified in thinking that we are alone in the world. There are however situations where this is not the case and

where different processes interfere with each other, either willingly or unwillingly. Here is my favourite example of such a case of interference[1].

Assume that we have three separate bytes in memory which contain the hour, minutes and seconds of the time of the day. There is a hardware device which produces an interrupt every second and this will cause the process that will update these three bytes to be woken up. Any process which wants to know the time, can access these three memory bytes, one after the other (we assume that our machine can address only one byte at a time). Now suppose that it is 10.59.59 and that a process has just read the first byte "10", when a clock interrupt occurs. As the process that updates the clock has a higher priority than the running process, the latter is suspended. The clock process now updates the time, setting it to 11.00.00. Control then returns to the first process which continues reading the next two bytes. The result is: 10.00.00; which is one hour wrong. What happened here is that *two processes access the same resource* — the three memory bytes — and that one or both of them can alter the contents. No harm would be done if both processes had *read-only* access to the shared resource.

The reader should note that the concept of a process has allowed us to speak about them as if they were really running simultaneously. We do not have to include in our reasoning the fact that there is a context switch and that complicated things are going on behind the scenes. We only have to be aware that access to shared resources must be protected, in order to avoid that another process accesses the same resource "simultaneously". On a multi-processor system "simultaneous" can really mean "at the same instant in time", on a uni-processor machine it really means "concurrently". The processor concept is equally valid for a uni- and a multi-processor machine.

The places in the program where a shared resource is accessed are so-called **critical regions**. We must avoid that two processes access simultaneously the resource and this can be done by ensuring that a process cannot enter a critical region when another process is already in a critical region where it accesses the same shared resource. The entrance to a critical region must be protected with a sort of a lock.

Two operations are defined on such a lock: *lock* and *unlock*. The lock operation tests the state of the lock and if it is unlocked, locks it. The test and the locking are done in a single **atomic** operation. If the lock is already locked, the lock operation will stop the process from entering the critical region. The unlock operation will simply clear a lock which was locked, and allow the other process access to the critical region again. That these operations must be *atomic* means that it must be impossible to interrupt them in the middle. Otherwise we would get into awkward situations again. If the lock operation would not be atomic, we could have a situation where process 1 inspects the lock and finds it open. If immediately after this, process 1 gets interrupted, before it had a chance to close the lock, process 2 could then also inspect the lock. It finds that it is open, sets it to closed, enters the critical region where it grabs the resource (a printer for

[1]The reader should be aware that the example describes a primitive situation; no modern operating system would allow this situation to occur.

instance) and starts using it. Some time later process 1 will run again, it will also close the lock and it will also grab the same printer and start using it. Remember that process 1 previously had found the lock to be open and it is unaware that process 2 has been running in the meantime!

The lock and unlock operations must therefore be completed before an interruption is allowed. This can be done —primitively— by disabling interrupts and then enabling them after the operation. No reasonable operating system would allow a normal user to tinker with the interrupts, so most machines have a *test-and-set* instruction. The *test-and-set* instruction tests a bit and sets it to "one" if it was "zero". If it was already "one" it is left unchanged. The result of the test (i.e. the state of the bit before the *test-and-set* instruction was executed is available in the processor status word and can be tested by a subsequent *branch* instruction. The *test-and-set* instruction is a single instruction; a hardware interrupt arriving during the execution of the instruction will be recognized only after the execution is complete. This guarantees the atomicity of a *test-and-set* operation.

What do we do after the *test-and-set* instruction? If the lock was open, you can safely enter the critical region. If, on the contrary, process A finds the lock closed, it should go to **sleep**. The operating system will then suspend the execution of process A and schedule another process to run, say C, or E. The process B, which had closed the lock in the first place, will also be running again at some instant and eventually will unlock the lock and **wakeup**[1] the sleeping process A. The system will then make process A runnable again.

Now suppose that process A gets interrupted immediately after doing its — unsuccessful — lock operation and before it could execute the *sleep()* call. Process B will at some stage open the lock and wakeup A. As A is not sleeping, this wakeup is simply lost. When A will run again, it will truly go to sleep, this time forever.

The solution to the problem was given in 1965 by Dijkstra, when he defined the **semaphore**. A semaphore can count the number of such "*lost wakeups*", without trying to wake up a process that is not sleeping. It can therefore only have a positive value, or "0". Two **atomic operations** are defined on a semaphore, which we will call **up** and **down**[2]. Once an operation on a semaphore is started, no other process can access the same semaphore. Thus **atomicity** of a semaphore operation is guaranteed. The work done for a *down* (and similarly for an *up*) operation must therefore be part of the operating system and not of a user process. The *down* operation checks the value of the semaphore. If it is greater than zero, it decrements the value and the calling process just continues execution. On the contrary, If the *down* operation finds that the semaphore value is zero, the calling process is put to sleep. The *up* operation on a semaphore increments its value. If one or more processes were sleeping, one of them is selected by the operating system. The selected process will then be allowed to run can now complete its *down*, which had failed earlier. Thus, if the semaphore was positive, it will simply

[1] Process B itself does not directly wakeup A, of course. The operating system takes care of doing it.

[2] Various other names are also used: post and signal, P and V (the original names given by Dijkstra), and possibly others. For mutexes, lock and unlock are often used.

be incremented, but if it was "0" — meaning that there are processes sleeping on it — its value will remain "0", but there will be one process less sleeping.

We have described the general form of a semaphore: the **counting semaphore**, which is used to solve **synchronization** problems, ensuring that certain events happen in the correct order. A **binary semaphore** can only take the values "0" or "1" and is particularly suited for solving problems of *mutual exclusion*, which explains its other name: **mutex**.

To illustrate the use of mutexes and counting semaphores we show an example of the **Producer-Consumer** problem. Suppose we have two collaborating processes: a *producer* which produces items and puts them in a *buffer* of finite size, and a *consumer* which takes items out of the buffer and consumes them. A data acquisition system which writes the collected data to tape is a good example of a producer-consumer problem. It is clear that the producer should stop producing when the buffer is full; likewise, the consumer should go to sleep when the buffer is empty. The consumer should wake up when there are again items in the buffer and the producer can start working again when some room in the buffer has been freed by the consumer.

```
#define N 100              /* number of slots in buffer */
typedef int semaphore;     /* this is NOT POSIX !!  */
semaphore mutex=1;         /* controls access to critical region */
semaphore empty=N;         /* counts empty buffer slots */
semaphore full=0;          /* counts full buffer slots*/

void producer(void)
{
  int item;
  while(TRUE) {            /* do forever (TRUE=1) */
      produce_item(&item); /* make something to put in buffer */
      down(&empty);        /* decrement empty count */
      down(&mutex);        /* enter critical region */
      enter_item(&item);   /* put new item in buffer */
      up(&mutex);          /* leave critical region */
      up(&full);           /* increment count of full slots */
  }
}


void consumer(void)
{
  int item;
  while(TRUE) {            /* do forever */
      down(&full);         /* decrement full count */
      down(&mutex);        /* enter critical region */
      remove_item(&item);  /* take item from buffer */
      up(&mutex);          /* leave critical region */
      up(&empty);          /* increment count of empty slots */
      consume_item(&item)  /* use the item */
  }
}
```

In order to obtain this synchronization between the two processes, two *counting semaphores* are used: *full* which is initialized to "0" and counts the buffer slots which are filled, and *empty*, initialized to the size of the buffer and which counts the empty slots. Access to the buffer, which is shared between the two processes, is protected by a *mutex*, initially "1" and thus allowing access. The example is taken from Andrew Tanenbaum's excellent book[1]. The reader should study carefully the listing of the *Producer-Consumer* problem on the previous page. He should be aware that the example is simplified: instead of two *processes* and a buffer structure in *shared memory*, the listing shows two functions, using global variables. Also the semaphores are not exactly what the standards prescribe. Using semaphores and mutexes remains a difficult thing: changing the order of two *down* operations in the listing below may result in chaos again.

1.4 What is wrong with Linux?

UNIX has the bad reputation of not being a real-time operating system. This needs some explanation. Time is an essential ingredient of a real-time system: the definition says that a real-time system must respond within a given time to an external stimulus. Theoretically, it is not possible to guarantee on a general UNIX time-sharing system that the response will occur within a specified time. Although in general the response will be available within a reasonable time, the load on the system cannot be predicted and unexpected delays may occur. It would be a bad idea to try and run a time-critical real-time application on an overloaded campus computer. Nevertheless, before discarding altogether the idea of using UNIX or Linux as the underlying operating system for a real-time application, we should have a critical look at what the requirements really are, to what extent they are satisfied by off-the-shelf Linux, and what can be done (or has been done already) to improve the situation.

The UNIX and Linux schedulers have been designed for **time-sharing** the CPU between a large number of users (or processes). It has been designed to give a *fair share of the resources*, in particular of CPU time, to all of these processes. The priorities of the various processes are therefore adjusted regularly in order to achieve this. For instance, the numerical analyst who runs CPU-intensive programs and does practically no I/O, will be penalized, to avoid that he absorbs all the CPU time.

Such a scheduling algorithm is not suitable for running a real-time application. If the operating system would decide that this particularly demanding application had consumed a sufficiently large portion of the available CPU time, it would lower its priority and the application might not be able anymore to meet its deadlines.

A real-time application must have **high priority** and — in order to be able

[1] Andrew S. Tanenbaum, Modern Operating Systems, Prentice Hall International Editions, 1992, ISBN 0-13-595752-4. The reader is encouraged to read the chapter on Interprocess Communication, which provides a much more detailed treatment of synchronization problems than is possible here.

to meet its deadlines — *must run whenever there is no runnable program with a higher priority*. In practice, the real-time process should have the highest priority, and it should keep this highest priority throughout its entire life[1]. Another scheduling algorithm is therefore required: a certain class of processes should be allocated permanently the highest priorities defined in the system. The normal scheduler of Linux did not have this feature, but *another scheduler*, designed for mixed time-sharing and real-time use is available and *is usually compiled into the kernel*.

Time being a precious resource for a real-time system, overheads imposed by the operating system should be avoided as much as possible. Some of the overheads can be avoided by careful design of the real-time program. For instance, knowing that forking a new process is a time-consuming business, all processes which the real-time application may need to run, should be forked and exec'ed (the *fork* and *exec* system calls will be illustrated in section 5) *during the initialization phase* of the application. Other overheads cannot be avoided so simply and need some adaptation or modification of the operating system.

Context switches may be very expensive in time, in particular when the code of the new process to be run is not yet available in memory and/or when room must be made in memory. All code and data of a real-time application should be **locked into memory**, so that this part of a context switch would not cause a loss of time. Locking everything into memory will also prevent *page faults* to happen, avoiding this way other *memory swapping* operations. Originally Linux did not have the possibility of locking processes into memory, but again, *memory locking* is now compiled into all recent kernels.

A further help in reducing the overheads due to context switches is to use so-called *light-weight processes* or **multi-threaded user processes**. Linux as such does not provide these, but *library implementations do exist* to implement the standard POSIX pthreads.

Other places where to watch for lurking losses of time are Input/Output operations. Normally, when a file is opened for writing, an initial block of disc sectors is allocated — usually 4096 bytes — and *inodes* and *directory entries* are updated. When the file grows beyond its allocated size, the relatively lengthy process of finding another free block of 4096 bytes and updating inodes and directory entries is repeated. A real-time system should be able to grab all the disc space it needs during initialization, so that these time losses may be avoided. Linux does not allow this at present.

All input and output in Linux is **synchronous**. This means that a process requesting an I/O operation will be *blocked until the operation is complete* (or an error is returned). Upon completion of the operation, the process becomes *runnable* again and it will effectively run when the scheduler decides so. However, "completion" of an output operation means only that the data have arrived in an output buffer, and there is no guarantee that the data have really been written out to tape or disc. When the process is only notified of completion of the I/O

[1] It would be wise to run a shell with an even higher priority, in order to be able to intervene when the real-time process runs out of hand. This shell would be sleeping, until it gets woken up by a keystroke.

operation when the data are really in their final destination, we have **synchronized I/O**, which may be a necessity for certain real-time problems. Linux does not spontaneously do *synchronized I/O*, but it *can be easily imposed by using sync or fsync*.

Asynchronous I/O may be another real-time requirement. It means that *the process* requesting the I/O operation *should not block* and wait for completion, *but continue processing* immediately after making the I/O system call. The standard device drivers of Linux do not work asynchronously, but the primitive system calls allow the option of continuing processing. A special purpose device driver could make use of this and thus do asynchronous I/O. The process will then be notified with an interrupt when the I/O operation has been completed.

The designer of a real-time system should of course also be aware that no standard device drivers exist for exotic[1] devices. They have to be written by the application programmer. In a standard UNIX system, such a new device driver must be compiled and linked into the kernel. Linux has a very nice feature: it allows to *dynamically load and link to the kernel so-called modules*, which can be — and very often are — device drivers.

We have shown before that it would be wise to divide a real-time system up into a set of processes, which can each care for their own business, without excessively interfering with each other. Nevertheless, some communication between processes may be needed. Old UNIX systems had only two interprocess communication mechanisms: **pipes** and **signals**. Signals have a very low information content, and only two user definable signals exist. System V UNIX added other IPC mechanisms: sets of **counting semaphores, message queues**, and **shared memory**. Most Linux kernels have the System V IPC features compiled in.

Probably no real-time system could live without a **real-time clock** and **interval timers**. They do exist in off-the-shelf systems, but the *resolution*, usually 1/50 th or 1/100 th of a second, may not be enough. The user-threads package can work with higher resolutions, if the hardware is adequate.

The IEEE has made a large effort to standardize the user interface to operating systems. The result of this effort has been the POSIX 1003.1a standard, which defines a set of system calls, and POSIX 1003.1b, which defines a standard set of Shell commands. Both were approved by IEEE and by ISO and thus gained international acceptance. Also **real-time extensions to operating systems** have been defined in the **POSIX 1003.1c-1994** standard, which has also been accepted by ISO. All the points discussed above are part of this POSIX 1003.1c standard, except for the **multiple threads and mutexes, which are defined in a later extension**. To the best of my knowledge, the so-called "**pthreads**" are now also an part of the international standard. Linux is POSIX 1003.1a and .1b compliant, although it may not have been officially certified.

[1] With exotic I really mean *very weird non-standard devices*. The list of devices supported by Linux is indeed incredibly long!

In summary, Linux used to be weak on the following, and may still be on a few items:

- **Mutexes**. A simple mutex did not exist in the original Linux kernel. The System V IPC semaphores can be used, although they are overkill, introducing a large overhead. Atomic bit operations are defined in *asm/bitops.h* and can be used more easily, but care should be exercised (danger of *priority inversion*). Mutexes are defined in the **pthreads** package. They will work between user threads inside a single process, and for some implementations also between threads and another process.

- **Interprocess Communication**. System V IPC is usually part of the Linux kernel and adds counting semaphores, message queues and shared memory to the usual mechanisms of pipes and of signals.

- **Scheduling**. A POSIX 1003.1c compliant scheduler for Linux exists and is part of the kernel in most *Linux distributions*.

- **Memory Locking**. Memory locking is part and parcel of the more recent Linux kernels (at least above 2.0.x and maybe earlier).

- **Multiple User Threads**. A few library implementations exist. The more recent Linux distributions have Leroy's Pthread library, which makes use of a particular feature of Linux: the "*clone*" system call. It is entirely compliant with the POSIX standard. You will soon get into close contact with it.

- **Synchronized I/O**. Can be obtained easily with *sync* and *fsync*.

- **Asynchronous I/O**. Not available in standard device drivers. Could be implemented for special purpose device drivers.

- **Pre-allocation of file space**. Not available to my knowledge.

- **Fine-grained real-time clocks and interval timers**. They are part of the available pthreads packages and could be used if the hardware is capable.

1.5 Creating Processes

Creating a new process from within another process is done with the *fork() system call*. *fork()* creates a new process, called the **child process**, *which is an exact copy of the original (parent) process*, including open files, file pointers etc. Before the *fork()* call there is only *one* process; when the *fork()* has finished its job, there are *two*. In order to deal with this situation, *fork()* **returns twice**. To the parent process it returns the **process identification (PID)** of the child process, which will allow the parent to communicate later with the child. To the child process it returns a 0. As the two processes are exact copies of each other, an *if* statement can determine if we are executing the child or the parent process.

There is not much use of a child process which is an exact copy of its parent, so the first thing the child has to do is to load into memory the program code that it should execute and then start execution at *main()*. A child is obviously too inexperienced to do this on its own, so there is a system call that does it for him: *execl()*. The entire operation of creating a new process therefore goes as follows:

```
/* here we have been doing things */
child=fork();          /* PID of new process --> child */
if(child){             /* here for parent process */
                       /*continue parent's business*/
    }
else {                 /* here for child process */
    execl("/home/boss/rtapp/toggle_rail_signal",\
     "toggle_rail_signal", N_sigs, NULL);
    perror("execl");   /* here in case of error */
    exit(1);
    }
/* here continues what the parent was doing */
```

execl() will do what was described above, so in our example it will load the executable file */home/boss/rtapp/toggle_rail_signal* and then start execution of the new process at *main(argc,argv)*. The other arguments of *execl()* are passed on to *main()*. *execl* is one of six variants of the exec system call: *execl*, *execv*, *execle*, *execve*, *execlp*, *execvp*. They differ in the way the arguments are passed to *main()*: **l** means that a list of arguments is passed, **v** indicates that a pointer to a vector of arguments is passed. **e** tells that environment pointer of the parent is passed and the letter **p** means that the *environment variable* **PATH** should be used to find the executable file.

This completes the creation of a new process. On a single CPU machine, one of the two processes may continue execution, the other will wait till the scheduler decides to run it. There is no guarantee that the parent will run before the child or vice versa.

The new process can *exit()* normally when it has done its job, or when it hits an error condition. The parent can *wait* for the child to finish and then find out the reason of the child's death by executing one of the following system calls:

```
pid_t wait(int *status); /* wait for any child to die*/
or: pid_t waitpid(pid_t which, int *status, int options)
            /* wait for child "which" to die */
```

These wait calls can be useful for doing some cleaning-up and to avoid leaving *zombies* behind. When the parent process exits, the system will do all the necessary clean-up, childs included.

We can now understand what the shell does when we type a command, such as *cp file1 dir*. The shell will parse the command line, and assume that the first word is the name of an executable file. It will then do a *fork()*, creating a copy of the shell, followed by an *execl()* or *execv()* which will load the new program, in our example the copy utility *cp*. The rest of the command line is passed on to *cp* as a

list or as a pointer to a vector. The shell then does a *wait()*. When an **&** had been appended to the command line, then the shell will *not* do a wait, but will continue execution after return from the *exec* call.

The following gives a more complete and rather realistic example of a *terminal server and a client*[1]. The reader is invited to study this example in detail.

The code for the **server** looks like:

```
#define     POSIX_C_SOURCE 199309

#include    <unistd.h>
#include    <stdio.h>
#include    <sys/types.h>
#include    <sys/wait.h>
#include    <signal.h>
#include    <errno.h>
#include    "app.h" /* local definitions */

main(int argc, char **argv)
{
    request_t r;
    pid_t terminated;
    int status;

    init_server();   /* set things up * /

    do {
        check_for_exited_children();
        r = await_request();  /*get some input*/
        service_request(r);  /*do what wanted*/
        send_reply(r);       /*tell we did it*/
    } while (r != NULL);

    shutdown_server();   /*tear things down*/
    exit(0);
}

void
service_request(request_t r)
{
    pid_t child;
    switch (r->r_op) {
        case OP_NEW:
            /* Create a new client process */
            child = fork();
```

[1] the example is taken from Bill O. Gallmeister, POSIX.4, Programming for the Real World, O'Reilly, 1995.

```
            if (child) {
                /* parent process */
                break;
            } else {
                /* child process */
                execlp("terminal","terminal \
                 application","/dev/com1",NULL);
                perror("execlp");
                exit(1);
            }
            break;
        default:
            printf("Bad op %d\n", r->r_op),
            break;
    }
    return;
}
```

The **terminal** end of the application looks like:

```
#include    <unistd.h>
#include    <stdio.h>
#include    <sys/types.h>
#include    <sys/wait.h>
#include    <signal.h>
#include    "app.h" /* local definitions */

char *myname;

main(int argc, char **argv)
{
    myname = argv[0];
    printf("Terminal \"%s\" here!", myname);
    while (1) {
        /* deal with the screen */
        /* await user input */
    }
    exit(0);
}
```

Presumably *request_t* is defined in *app.h* as a pointer to a structure. *await_request()* is a function which sleeps until a service request arrives from a terminal. The operations performed by the other functions:
init_server, *service_request()*, *check_for_exited_children()*, *send_reply()* and *shutdown_server()* are implied by their names.

1.6 Interprocess Communication

In the case where we have a real-time application with a number of processes running concurrently, it would be a normal situation when some of these processes need to communicate between them. We said already that the classical UNIX system only knows pipes and signals as communication mechanisms. Interprocess communication, suitable for real-time applications is an essential part of the POSIX standard, which adds a number of mechanisms to the minimal UNIX set. In the following we will briefly describe the various IPC mechanisms and how they can be invoked. We will follow as much as possible the POSIX standard, except where the facilities are not implemented in Linux. In that case we will describe the mechanism Linux makes available.

1.6.1 UNIX and POSIX 1003.1a Signals

The old signal facility of UNIX is rather limited, but it is available on every implementation of UNIX or one of its clones. Originally, signals were used to *kill* another process. Therefore, for historical reasons, the system call by which a process can send a signal to another process is called *kill()*. There is a set of signals, each identified by a number (they are defined in <*signal.h*>), and the complete system call for sending a signal to a process is:
`kill(pid_t pid, int signal);`
The integer *signal* is usually specified symbolically: *SIGINT*, *SIGALRM* or *SIGKILL*, etc., as defined in <*signal.h*>. *pid* is the process identification of the process to which the signal shall be sent. If this receiving process has not been set up to **intercept signals**, its *execution will simply be terminated by any signal sent to it*. The receiving process can however be set up to *intercept certain signals* and to perform certain actions upon reception of such an intercepted signal. Certain signals cannot be intercepted, they are just killers: SIGINT, SIGKILL are examples. In order *to intercept a signal*, the receiving process must have *set up a signal handler* and notified this to the operating system with the *sigaction() system call*. The following is an example of how this can be done:

A *structure sigaction* (not to be confounded with the system call of the same name!) is defined as follows:

```
struct sigaction {
    void (*sa_handler)();
    sigset_t sa_mask;
    int sa_flags;
void(*sa_sigaction)(int,siginfo_t *,void *); };
```

This structure encapsulates the action to be taken on receipt of a signal.

The following is a program that shall exit gracefully when it receives the signal *SIGUSR1*. The function *terminate_normally()* is the **signal handler**. The administrative things are accomplished by defining the elements of the structure and then calling *sigaction()* to get the signal handler registered by the operating system.

```
void
terminate_normally(int signo)
{
    /* Exit gracefully */
    exit(0);
}

main(int argc, char **argv)
{
    struct sigaction sa;
    sa.sa_handler = terminate_normally;
    sigemptyset(&sa.sa_mask);
    sa.sa_flags = 0;
    if (sigaction(SIGUSR1, &sa, NULL)) {
        perror("sigaction");
        exit(1);
    }
 ...
}
```

The operating system itself may generate signals, for instance as the result of machine exceptions: *floating point exception, page fault*, etc. Signals may also be generated by something which happens *asynchronously* with the process itself. The signals then *aim at interrupting the process*: I/O completion, timer expiration, receipt of a message on an empty message queue, or typing CTRL-C or CTRL-Z on the keyboard. Signals can also be sent from one user process to another.

The structure *sigaction* does not only contain the information needed to register the signal handler with the operating system (in the process descriptor), but it also contains information on what the receiving process should do when it receives the registered signal. It can do one of three things with the signal:

— it can block the signal for some time and later unblock it.
— it can ignore the signal, pretending that nothing has happened.
— it can handle the signal, by executing the signal handler.

The POSIX.1 signals, described so far, have some serious limitations:

— there is a lack of signals for use by a user application (there are only two: SIGUSR1 and SIGUSR2).
— signals are not queued. If a second signal is sent to a process before the first one could be handled, the first one is simply and irrevocably lost.
— signals do not carry any information, except for the number of the signal.
— and, last but not least, signals are sent and received asynchronously. This means in fact that a process may receive a signal at any time, for instance also when it is updating some sensitive data-structures. If the signal handler will also do something with these same data-structures, you may be in deep trouble. In other words, when you write your program, you must always keep in mind that you may receive a signal exactly at the point where your pencil is.

Linux is compliant with this POSIX 1003.1a definition of signals.

1.6.2 POSIX 1003.1c signals

From the description above, we have seen that the POSIX 1003.1a signals are a rather complicated business (in UNIX jargon this is called flexibility). The POSIX 1003.1c extensions to the signal mechanism introduces even more flexibility. POSIX 1003.1c really defines an entirely *new set of signals*, which can peacefully co-exist with the old signals of POSIX 003.1a1. The historical name *kill()* is replaced by the more expressive *sigqueue()*.

The main improvements are:
– a far larger number of user-definable signals.
– signals can be queued; old untreated signals are therefore not lost.
– signals are delivered in a fixed order.
– the signal carries an additional integer, which can be used to transmit more information than just the signal number.

POSIX 1003.1c signals can be sent *automatically* as a result of *timer expiration, arrival of a message on an empty queue*, or *by the completion of an asynchronous I/O operation*. Unfortunately, the POSIX 1003.1c signals may not be part of Linux, so we will not dwell on them any further.

1.6.3 pipes and FIFOs

Probably one of the oldest interprocess communication mechanisms is the **pipe**. Through a pipe, the *standard output* of a program is pumped into the *standard input* of another program. A pipe is usually set up by a shell, when the pipe symbol (|) is typed between the names of two commands. The data flowing through the pipe is lost when the two processes cease to exist. For a **named pipe**, or **FIFO** *(First In, First Out)*, the data remains stored in a file. The *named pipe* has a name in the filesystem and its data can therefore be accessed by any other process in the system, provided it has the necessary permissions.

A running process can set up a pipe to communicate with another process. The communication is uni-directional. If duplex communication is needed, two pipes must be set up: one for each direction of communication. The two “ends of a pipe” are nothing else than file descriptors: one process writes into one of these files, the other reads from the other.

Setting up a pipe between two processes is not a terribly straightforward operation. It starts off by making the *pipe()* system call. This creates *two file descriptors*, if the calling process still has file descriptors available. One of these descriptors (in fact the second one) concerns the end of the pipe where we will write, the other descriptor (the first one) is attached to the opposite end, where we will read from the pipe. If we now create another process, this newly created process will inherit these two file descriptors. We now must make sure that both parent and child processes can find the file descriptors for the pipe ends. The *dup2* system call will in fact do this, by duplicating the “abstract” file descriptors *pipe_ends[0]* and *pipe_ends[1]* into well-known ones. *dup2* copies a file descriptor into the first available one, so we should close first the files where we want the

pipe to connect (usually *standard out* for the process connected to the writing end and *standard in* for the process which will read from the pipe). Here is a *skeleton* program for doing this in the case of a *terminal server*, which *forks off a terminal process to display messages from the server*:

```
/* First create a new client */
if (pipe(pipe_ends) < 0) {
    perror("pipe");
    exit(1);
}

global_child=child=fork();
if (child) {
    /*here for parent process*/
    do_something();
}
else {
    /*here for the child*/
    /* pipe ends will be 0 and 1 (stdin and stdout) */
    (void)close(0);
    (void)close(1);
    if (dup2(pipe_ends[0], 0) < 0)
        perror("dup2");
    if (dup2(pipe_ends[1], 1) < 0)
        perror("dup2");
    (void)close(pipe_ends[0]);
    (void)close(pipe_ends[1]);
    execlp(CHILD_PROCESS, CHILD_PROCESS, "/dev/com1", NULL);
    perror("execlp");
    exit(1);
}
```

The terminal process, created as the child could look:

```
#include <fcntl.h>
char buf[MAXBYTES]
/* pipe should not block, to avoid waiting for input */
if(fcntl(channel_from_server, F_SETFL, O_NONBLOCK) < 0){
    perror("fcntl");
    exit(2);
}
while (1) {
    /* Put messages on the screen */
    /* check for input from the server */
    nbytes = read(channel_from_server, buf, MAXBYTES);
    if (nbytes < 0) && (errno != EAGAIN))
        perror("read");
    else if (nbytes > 0) {
        printf("Message from the Server:  \"%s\"\n", buf);
}
...
```

In this example[1], the server process simply writes to the write end of the pipe (which has become stdout) and the child reads from the other end, which has been transformed by *dup2* into stdin. To set up a communication channel in the other direction as well, the whole process must be repeated, inverting the roles of the server and the terminal client (the first becomes the reader, and the second the writer) and using two other file descriptors (for instance 3 and 4 if they are still free). Note that the *dup* calls must be made before the child does its *exec* call, otherwise, the file descriptors for the two pipe ends would be lost.

The use of named pipes is simpler: the *FIFO* exists in the file system and any process wanting to access the file can just open it. One process should open the *FIFO* for reading, the other for writing. A FIFO is created with the POSIX 1003.1a *mkfifo()* system call.

1.6.4 Message Queues

When we have compiled the *System V IPC facilities* into the Linux kernel, we have **message queues** available, which however do not conform to the POSIX 1003.1c standard. We will nevertheless describe them briefly, as they are the only ones we have at present.

In system V the **message resource** is described by a *struct msqid_ds*, which is allocated and initialized when the resource is created. It contains the permissions, a pointer to the last and the first message in the queue, the number of messages in the queue, who last sent and who last received a message, etc. The messages itself are contained in:

```
struct msgbuf {
       long mtype;
       char mtext[l]; }
```

To set up a message queue, the creator process executes a *msgget* system call:

```
msqid = msgget(key_t key, int msgflg);
```

The *msqid* is a unique identification of the particular message queue which ensures that messages are delivered to the correct destination. The exact role of the key is complicated; in most cases the key can be chosen to be *IPC_PRIVATE*. The use of *IPC_PRIVATE* will create a new message queue if none exist already. If you want to do unusual things or make full use of the built-in *flexibility*, you may fabricate your own key with the *ftok(char* pathname, char proj)* library call and play with *msgflg*.

A process wanting to receive messages on this queue must also perform a *msgget* call, in order to obtain the *msqid*.

A message is sent by executing:

```
int msgsnd(int msqid,struct msgbuf *msgp,int msgsz, \
int msgflg);
```

[1]Which was also taken from Gallmeister's book.

and similarly a message is received by:
```
int msgrcv(int msqid, struct msgbuf *msgp, int msgsz, \
long msgtyp, int msgflg);
```
msgtyp is used as follows:

if msgtyp = 0 : get first message on the queue,
> 0 : get first message of matching type,
< 0 : get message with smallest type
which is $\leq$abs(msgtyp).

Finally, the *msgctl* calls allow you to get the status of a queue, modify its size, or destroy the queue entirely.

The message queue can be empty. If a message is sent to an empty queue, the process reading messages from the queue is woken up. Similarly, when the queue is full, a writer trying to send a message will be blocked. As soon as a message is read from the queue, creating space, the writer process is woken up.

1.6.5 Counting Semaphores

System V **semaphore arrays** are an **oddity**. The *semget* call allocates an **array of counting semaphores**. Presumably, and hopefully, the array may be of length 1. You also specify operations to be performed on a series of members of the array. The operations are only performed if they will **all succeed!**

Counting semaphores can be useful in **producer-consumer problems**, where the producer puts items in a buffer and the consumer takes items away. Two counting semaphores keep track of the number of items in the buffer and allow to "gracefully" handle the *buffer empty* and *buffer full* situations.

Producer-consumer situations can easily arise in a real-time application: the *producer collects data from measuring devices*, the *consumer writes the data to a storage device* (disk or tape).

Another example is a large paying car park: There is one *counting semaphore* which is initialized to the total number of places in the car park. A *separate process is associated* with *each entrance or exit gate*. The process at an entrance gate will **do a wait on the semaphore, e.g. decrement it**. If the result is greater than zero, the process will continue, issue a ticket with the time of entrance, and open the gate. It closes the gate as soon as it has detected the passage of the car. If the *value of the counting semaphore* is zero when the decrement operation is tried, the process is **blocked** and added to the pile of blocked processes. This is just what is needed: the car park is full and the car will have to wait, so no ticket is issued, etc.

The processes at the exit gates do the contrary: after having checked the ticket, they open the gate and then do a *post or increment* operation on the semaphore, effectively indicating that one more place has become free. This operation will always succeed.

The *System V counting semaphore mechanism* is rather similar to the message queue business: You create a semaphore (array) as follows:
```
int semid = semget(key_t key, int nsems, int semflg);
```

The key IPC_PRIVATE behaves as before. All processes wanting to use the semaphore must execute this *semget* call. You can then operate on the semaphore:
```
int semop(int semid, struct sembuf *sops, \
unsigned nsops);
```
(here is the oddity, you do nsops operations on nsops members of the array; the operations are specified in an array of *struct sembuf*). This structure is defined as:

```
struct sembuf
       ushort sem_num;  /*index in array*/
       short sem_op;    /*operation*/
       short sem_flg    /*operation flags*/
```

Two kinds of operations can result in the process getting blocked:
i) If *sem_op* is 0 and *semval* is non-zero, the process *sleeps* on a queue, waiting for *semval* to become zero, or returns with error *EAGAIN* if either of *(IPC_NOWAIT | sem_flg)* are true.
ii) If *(sem_op < 0)* and *(semval + sem_op < 0)*, the process either sleeps on a queue waiting for *semval* to increase, or returns with error *EAGAIN* if *(sem_flg & IPC_NOWAIT)* is true.

Atomicity of the semaphore operations is guaranteed, because the mechanism is embedded in the kernel. The kernel will not allow two processes to simultaneously use the kernel services. In other words, a system call will be entirely finished before a context switch takes place.

Note: If you want to use a semaphore which takes only the values 0 or 1 (for instance for mutual exclusion), you are better off by using the atomic bit operations, defined in <asm-i386/bitops.h>: *test_bit, set_bit and clear_bit*.

1.6.6 Shared Memory

Shared Memory is exactly what its name says: two or more processes **access the same area of physical memory**. This segment of physical memory is **mapped into** two or more **virtual memory spaces**.

Shared Memory is considered a low-level facility, because the shared segment **does not benefit from the protection** the operating system normally provides. To compensate for this disadvantage, **shared memory is the fastest IPC mechanism**. The processes can read and write shared memory, without *any system call being necessary*. The *user himself must provide the necessary protection*, to avoid that two processes "simultaneously" access the shared memory. This can be obtained with a **binary semaphore** or **mutex**.

A *mutex* can be simulated by performing the *set_bit(int nr, void * addr)* call, which sets the desired bit *nr* and returns the old value of the bit. The short integer on which this operation is performed must also reside in shared memory, in order to be accessible by both processes.

The shared memory facility available in Linux comes from System V, and is may therefore be **not** conforming to POSIX.1c. The related system calls are similar to the System V calls we have already seen:

There is, of course, a *shared memory descriptor*,
`struct shmid_ds`.
Shared memory is allocated with the system call:
`shmid = shmget(key_t key, int size, int shmflg);`
The *size* is in bytes and should preferably correspond to a multiple of the page size (4096 bytes). All processes wanting to make use of the shared memory segment must make a *shmget* call, with the same key.
Once the memory has been allocated, you map it into the virtual memory space of your process with:

```
char *virt_addr;
virt_addr = shmat(int shmid, char *shmaddr, int shmflg);
```

shmaddr is the requested attach address:
if it is 0, the system finds an unmapped region;
if it is non-zero, then the value must be *page-aligned*.
By setting *shmflg = SHM_RDONLY* you can request to attach the segment *read-only*.

You can get rid of a shared memory segment by:
`int shmdt(char *virt_addr);`
Finally, there is again the *shmctl* call, which you may use to get the status, or also to destroy the segment (a shared segment will only be destroyed after all users have *detached* themselves).

If you are using shared memory, and you need *malloc* as well, you should `malloc` a large chunk of memory first, before you attach the shared memory segment. Otherwise *malloc* may interfere with the shared memory.

A word about the Linux implementation of the System V IPC mechanisms is in order. All System V system calls described above make use of a single Linux system call: ipc(). A library of the system V IPC calls is available, which maps each call and its parameters into the Linux ipc() call. An example is:

```
int semget (key_t key, int nsems, int semflg)
{
    return ipc (SEMGET, key, nsems, semflg, NULL);
}
```

The constants are defined in <linux/ipc.h>

1.7 Scheduling

The original scheduling algorithm of Linux aimed at giving a **fair share of the resources** to each user. It therefore was a typical **time-sharing scheduler**. A time-sharing scheduler is based on priorities, like any other type of scheduler,

but the system keeps changing the priorities to attain its aim of being fair to everyone.

For *time-critical real-time applications* you want another sort of scheduler. You need a **high priority for the most critical real-time processes**, and a *scheduler* which will run such a high priority process whenever **no process with higher priority is runnable**[1].

Less critical processes of the real-time application can run at *lower priorities* and other user jobs could also be fitted in at priorities below.

SVR6 (System V, Release 6) has a scheduler that does both time-sharing and real-time scheduling, depending on the priority assigned to a process. Critical processes run at priorities between, say, 0 and 50, and benefit from the priority scheduling. Other jobs run at lower priorities and have to accept the time-sharing scheduler. This aspect of System V has not been ported to Linux.

A POSIX.1c compliant scheduler **has** been ported to Linux. In order to make use of it, you must make *patches to the kernel code* and recompile the kernel together with this POSIX.1c scheduler. At the time these notes were prepared, we had not yet had a chance to try it.

The advantage of a POSIX.1c scheduler is, of course, that your application program will be portable between different platforms.

What does a POSIX.1c scheduler do? Here is what it provides[2]:

```
#include          <unistd.h>
#ifdef _POSIX_PRIORITY_SCHEDULING
#include          <sched.h>
int i, policy;
struct sched_param *scheduling_parameters;
pid_t pid;

sched_setscheduler(pid_t pid, int policy, \
    struct sched_param *scheduling_parameters);
int sched_getscheduler(pid_t pid);
int sched_getparam(pid_t pid, \
    struct sched_param *scheduling_parameters);
int sched_setparam(pid_t pid, \
    struct sched_param *scheduling_parameters);
int sched_yield(void);
int sched_get_priority_min(int);
int sched_get_priority_max(int);
#endif _POSIX_PRIORITY_SCHEDULING
```

You see that you define a *scheduling "policy"*. You have a choice:

SCHED_FIFO: pre-emptive, priority-based scheduling,
SCHED_RR: pre-emptive, priority-based with *time quanta*,
SCHED_OTHER: implementation dependent scheduler.

[1] Remember that you need a sleeping shell at a still higher priority.
[2] Again from Gallmeister's book.

With the first choice, the process *will run* until it gets blocked for one reason or another, or *until a higher priority process becomes runnable*. The second policy adds a **time quantum**: a process running under this scheduling policy will only run for a certain duration of time. Then it goes back to the end of the queue for its priority (each priority level has its own queue). Thus, at a given priority level, all processes in that level are scheduled **round-robin**. In future, **deadline scheduling** will probably have to be added as another choice.

There is a range of priorities for the *FIFO* scheduler and another range for the *RR* scheduler.

After a *fork()*, the child process *inherits the scheduling policy and the priority* of the parent process. If the priority of the child then gets increased above the priority of the running process, the latter is *immediately pre-empted*, even before the return from the *sched_setparam* call! So be careful, you may seriously harm yourself.

On the other hand, you may *"yield"* the processor to another process. You cannot really be sure which process this is going to be. As a matter of fact, the only thing *yield* does, is to put your process at the end of the queue at your particular priority level.

All this is nice, but we are still *stuck with the fact that the kernel itself cannot be pre-empted*. This is usually not too much of a problem. Most of the system calls will take only a short time to execute.

Usually, the system calls that may take a considerable time (such as certain I/O related calls), should be relegated — as far as possible — to those tasks that run at a lower priority level. Also some common sense will help: it is much faster to write once 512 bytes to disk than to write 512 times a single byte!

Other system calls do take a long time. *fork* and *exec* for example. You should therefore **create** all necessary processes during the **initialization phase** of your application. Let the processes that you only need sporadically just *sleep* for most of the day.

1.8 Timers

You may want to arrange for certain things to happen at **certain times**, or a given **time interval after** something else happened. So you will nearly always have the need for a **timer** and/or an **interval timer**.

Standard UNIX (and Linux) has a **real-time clock**. It counts the number of seconds since 00:00 a.m. January 1, 1970. (called the *Epoch*). You get its value with the *time()* function:

```
#include <time.h>
time_t time(time_t *the_time_now);
```

You can also call *time* with a NULL pointer.

Linux also has the *gettimeofday* call, which stores the time in a structure:

```
struct timeval {
    time_t tv_sec   /* seconds */
    time_t tv_usec }    /* microseconds */
```

`gettimeofday` returns a 0 or -1 (success, failure respectively).

You can make things happen **after a certain time interval** with *sleep*:

```
unsigned int sleep(unsigned int n_seconds);
```

The process which executes this call will be stopped and resumed *after n_seconds* have passed. The resolution is very crude! As a matter of fact, many real-time systems would need a resolution of milliseconds and, in extreme cases, even microseconds.

To overcome this drawback, Linux has also **interval timers**. each process has three of them:

```
#include      <sys/time.h>

int setitimer(int which_timer, \
    const struct itimerval *new_itimer_value, \
    struct itimerval *old_itimer_value);
int getitimer(int which_timer, \
    struct itimerval *current_itimer_value);
```

The first argument, *which_timer*, has one of three values:
ITIMER_REAL, ITIMER_VIRTUAL and *ITIMER_PROF*. *setitimer()* sets a new value of the interval timer and returns the old value in *old_timer_value*.

When a timer **expires, it delivers a signal**: *SIGALRM, SIGVTALRM* and *SIGPROF* respectively. The calls make use of a structure:

```
struct itimerval {
    struct timeval it_val    /* initial value */
    struct timeval it_interval }    /* interval */
```

The *ITIMER_REAL* measures the time on the “wall clock” and therefore includes the time used by other processes.
ITIMER_VIRTUAL measures the time spent in the user process which set up the timer, whereas *ITIMER_PROF* counts the time spent in the user process **and** in the kernel on behalf of the user process. It is thus very useful for *profiling*.

The resolution of these interval timers is given by the constant *HZ*, defined in *<sys/param.h>*. On Linux machines, *HZ=100*, so the resolution of the interval timers is 10000 microseconds.

POSIX.1c extends the timer facilities to a number of implementation defined clocks, which may have different characteristics. Timers and intervals can be specified in nanoseconds.

1.9 Memory Locking

As we already pointed out before, the real-time processes – at least the critical ones – should be **locked into memory**. Otherwise you could have the very unfortunate situation that your essential task has been swapped out, just before it becomes runnable again. **Faulting** a number of pages of code back into memory may add an intolerable overhead.

Remember also that *infrequently used pages may be swapped out* by the system, without any warning. Faulting them back in again may make you miss a **deadline**. Thus, not only the *program code*, but also the *data and stack pages* should be locked into memory.

A POSIX.1c conformant memory locking mechanism is available for Linux. Unfortunately, we have not yet been able to test it. It does the following:

```
#include       <unistd.h>
#ifdef _POSIX_MEMLOCK
#include       <sys/mman.h>
int mlockall(int flags);
int munlockall(void);
#endif /* _POSIX_MEMLOCK */
```

mlockall will lock **all** your memory, e.g. *program, data, heap, stack and also shared libraries*. You may choose, by specifying the flags, to lock the space you occupy at present, but also what you will occupy in future.

Instead of locking everything, you may also lock parts:

```
#include       <unistd.h>
#ifdef _POSIX_MEMLOCK_RANGE
#include       <sys/mman.h>

int mlock(void *address, size_t length);
int munlock(void *address, size_t length);
#endif /* _POSIX_MEMLOCK_RANGE */
```

Finally, you may want to lock just a few essential functions: a signal handler or an interrupt handler, for instance. You should not do this from within the interrupt handler, but from a separate function:

```
void intr_handler()
{
    ...  /* do your work here */
}
void right_after_intr_handler()
{
        /* this function serves to get an address */
        /* associated with the end of intr_handler() */
}
```

```
void intr_handler_init()
{
    ...
    i = mlock(ROUND_DOWNTO_PAGE(intr_handler),\
      ROUND_UPTO_PAGE(right_after_intr_handler - \
      intr_handler));
}
```

The function *right_after_intr_handler()* does nothing. It serves only to get an address associated with the *end* of the interrupt handler. This is needed to calculate the argument *length* for the *mlock()* call.

1.10 Multiple User Threads

All we have seen so far happened at the *process* level and *kernel intervention* was needed for every coordinating action between processes. The overall picture has become quite complicated and a programmer must master many details or else he runs into trouble.

Is there not another solution, where the user has more direct control over what is going on? Fortunately, there is: **multiple user threads**. POSIX.4a (or POSIX.1c if you prefer) standardizes the **API (Application Programmer's Interface)** for multiple threads.

Threads are independent flows of control *inside a single process*. Each *thread* has its *own thread structure* — comparable to a *process descriptor* —, its *own stack* and its *own program counter*. All the rest, i.e. *program code, heap storage* and *global data*, is **shared** between the threads. Two or more threads may well execute the same function simultaneously. The services needed to *create threads, schedule their execution, communicate and synchronize between threads* are provided by the **threads library** and *run in user space*. For the kernel exists only the *process*; what happens inside this process is invisible to the kernel.

Lightweight Processes, as in Solaris or SunOS 4.x, are somewhere mid-way: a small part of the process structure has been split off and can be replicated for several LWPs, all continuing to be part of the same process, using the same memory map, file descriptors, etc. The split-off part is still a kernel structure, but the kernel can now make rapid context switches between LWPs, because only a small part of the complete process structure is affected. Inside a LWP, multiple threads may be present.

Multiple threads offer a solution to programming which has a number of advantages. The model is particularly well suited to *Shared-memory Multiple Processors*, where the code, common to all threads, is executed on different processors, one or more threads per processor. Also for real-time applications on uniprocessors, threads have advantages. In the first place, the *fastest, easiest intertask communication mechanism, — shared memory —* is there for *free*!

There are other advantages as well. The **responsiveness** of the process may increase, because when one thread is *blocked*, waiting for an event, the other can continue execution. The fact that threads offer a sort of "do-it-yourself" solution

makes the user have a better grasp of what he is doing and thus he can produce better structured programs. *Communication* and *synchronization* between threads is easier, more transparent and faster than between processes. Each thread conserves its ability to communicate with another process, but it is wise to concentrate all *inter-process communication* within a single thread.

Multiple threads will in general lead to performance improvements on shared memory multiprocessors, but on a uniprocessor one should not expect miracles. Nevertheless, the fact that there is **less overhead** and that some threads may **block while others continue**, will be felt in the **performance**.

It sounds as if we just discovered a gold mine. Well . . . , there are a few things which obscure the picture somewhat. For threads to be usable with no danger, the **library functions** our program uses must be **threads-safe**. That is, they must be *re-entrant*. Unfortunately, many libraries contain functions which modify global variables and therefore are **not** re-entrant. For the same reason, your threaded program must be re-entrant, so it has to be compiled with *_REENTRANT* defined. In addition, for a real-time application, you still need at least a few facilities from the operating system: *memory locking and real-time priority scheduling*[1].

Threads can be implemented as a *library of user functions*. The standard set of functions is defined in POSIX.1c, but other implementations also exist. The package we are using[2] implements the **POSIX.1c pthreads**. There are some 50 service requests defined. They are briefly described in Annex III and in more detail in the man pages. We will illustrate only a few of them, the most important ones.

Pthreads defines functions for *Thread Management, Mutexes, Condition Variables, Signal Management, Thread Specific Data* and *Thread Cancellation*. Threads, mutexes and condition variables have *attributes*, which can be modified and which will change their behaviour. Not all options defined by the various attributes need to be implemented. <*pthread.h*> defines eight data types:

Type	**Description**
`pthread_attr_t`	`Thread attribute`
`pthread_mutexattr_t`	`Mutex attribute`
`pthread_condattr_t`	`Condition variable attribute`
`pthread_mutex_t`	`Mutual exclusion lock (mutex)`
`pthread_cond_t`	`Condition variable`
`pthread_t`	`Thread ID`
`pthread_once_t`	`Once-only execution`
`pthread_key_t`	`Thread specific data key`

[1] Alternatively, you run on a dedicated machine, where you have killed all daemons, so that your application is the only active process in the system.

[2] Xavier Leroy's implementation, called **LinuxThreads**, which is part of many recent Linux distributions (Xavier.Leroy@inria.fr).

Attributes can be *set or retrieved* with calls of the following type:

```
int pthread_attr_setschedpolicy( pthread_attr_t *attr, int newvalue);
or:
int pthread_mutexattr_getprotocol( pthread_mutexattr_t *attr, *protocol);
```

See Annex III for the complete list. The *scheduling policy* can be one of: `SCHED_FIFO`, `SCHED_RR` and `SCHED_OTHER`, as for the POSIX.1c standard. The scheduling parameters can also be set and retrieved.

When the process is forked, *main(argc, argv)* is entered. In the main program you may then create threads. Each thread is a function, or a sequence of functions. At thread creation, the *entry point* must be specified:

```
int pthread_create( pthread_t *thread, \
          const pthread_attr_t *attr, \
          void *(*entry)(void *), \
          void *arg );

   void pthread_exit( void *status );
```

does what is expected from it. It should be noted that *NULL* may often be used to substitute an argument in the function call. This is notably the case for `pthread_attr_t *attr` and `void *status` above.

An important function is:

```
int pthread_join( pthread_t thread, void **status );
```

When this primitive is called by the running thread, its execution will be suspended until the target `thread` terminates. If it has already terminated, execution of the calling thread continues. *pthread_join()* is therefore an important mechanism for synchronizing between threads. So-called *detached threads* cannot be joined. You specify at creation time or at run time if the thread has to be detached or not.

Mutexes can have as the *pshared* attribute `PTHREAD_PROCESS_SHARED` or `PTHREAD_PROCESS_PRIVATE`, meaning that the mutex can be accessed also by other processes or that it is private to our process. Private mutexes are defined in all implementations, shared mutexes are an option. The two usual operations on a mutex are:

```
int pthread_mutex_lock( pthread_mutex_t *mutex );
```

and

```
int pthread_mutex_unlock( pthread_mutex_t *mutex );
```

but you can also try if a mutex is locked and continue execution, whatever the result:

```
int pthread_mutex_trylock( pthread_mutex_t *mutex );
```

All memory occupied by the process is shared among the various threads, which we said was an important advantage of threads. Nevertheless, sometimes a thread needs to protect its data against attacks from other threads. For this reason a few primitives which allow to create and manipulate *thread specific data* are defined. For details see the man pages.

We have not yet met **condition variables**, which are another feature of pthreads. Condition variables are always associated with a *mutex*. A condition variable is used to *signal* a thread that a particular condition has been satisfied in another thread. The first thread — the one receiving the signal — will then be allowed to proceed if it had blocked on the condition variable (CV). It works as follows:

Thread 1	**Thread 2**
lock the mutex	
test the condition	
FALSE! *unlock mutex*	
sleep on CV	
	lock the mutex
	change the condition
	signal thread 1
	unlock mutex
lock mutex	
test condition again	
TRUE! *do the job*	
unlock mutex	

Translated into code, this becomes:

Thread 1	**Thread 2**
`pthread_mutex_lock(&m);`	
`while (!my_condition) {`	
`while (pthread_cond_wait(&c, &m) != 0) { ;`	
	`pthread_mutex_lock(&m);`
	`my_condition = TRUE ;`
	`pthread_cond_signal(&c);`
	`pthread_mutex_unlock(&m);`
`do_thing();`	
`}`	
`}`	
`pthread_mutex_unlock(&m);`	

Note that *pthread_cond_wait()* will automatically free the mutex for you and your thread will go to sleep on the condition variable.

pthreads is really a subject in itself and our quick review has been very superficial. Threads are well suited for implementing **Server-Client problems**. Due to the shared memory, the communication between the *server* and the — possibly many — *clients* is easy.

We close this section with a complete code example[1]. The example concerns an *Automatic Teller Machine*, e.g. one of those machines that distribute banknotes.

[1]This and the following example are from *B. Nichols et.al., Pthreads Programming* See Bibliography, item ii)

The main program, which is the **server**, receives requests over a communication line from ATMs scattered all over town. For each request received, the server spawns a *worker* or *client thread* which undertakes the actions necessary to satisfy the request. This example mainly illustrates the creation of several threads.

```
typedef struct workorder {
        int conn;
char req_buf[COMM_BUF_SIZE];
} workorder_t;

main(int argc, char **argv)
{
  workorder_t *workorderp;
  pthread_t   *worker_threadp;
  int  conn, trans_id;

  atm_server_init(argc, argv);

  for(;;) {
    /*** Wait for a request ***/
    workorderp = (workorder_t *)malloc(sizeof(workorder_t));
    server_comm_get_request(&workorderp->conn,
            &workorderp->req_buf);

    sscanf(workorderp->req_buf, "%d", &trans_id);
    if (trans_id == SHUTDOWN) {
          . . .
      break;
    }

    /*** Spawn a thread to process this request ***/
    worker_threadp = (pthread_t *)malloc(sizeof(pthread_t));
    pthread_create(worker_threadp, NULL, process_request,
          (void *)workorderp);

    pthread_detach(*worker_threadp);
    free(worker_threadp);
  }
  server_comm_shutdown();
}
```

The *worker thread* (the **client**) looks as follows:

```
void process_request(workorder_t *workorderp)
{
  char resp_buf[COMM_BUF_SIZE];
  int trans_id;
  sscanf(workorderp->req_buf, "%d", &trans_id);

  switch(trans_id) {
    case WITHDRAW_TRANS:
```

```
         withdraw(workorderp->req_buf, resp_buf);
 break;

    case BALANCE_TRANS:
         balance(workorderp->req_buf, resp_buf);
 break;
 .
 .
    default:
         handle_bad_trans_id(workorderp->req_buf, resp_buf);
  }

  server_comm_send_response(workorderp->conn, resp_buf);
  free(workorderp);
}
```

There are two points to note in this example. The first concerns the passing of arguments to a child thread. The standard allows a single argument only. Encapsulating several arguments in a single structure and passing a pointer to this structure to the child is a way to program around the restriction. The second point is a subtle one and concerns the use of *malloc*. Using static storage for the workorder does not work: for every newly created thread the workorder would be overwritten and most threads would work with a corrupted workorder.

The following example, taken from the same source, illustrates the use of a *mutex* and a *condition variable*. Two of the threads created in this example simply increment a counter and check if it has reached a limit value. In that case they *signal the condition variable*. The third thread waits on the condition variable and prints its value. The main thread will exit when all three threads it created have "*joined*".

```
#include <pthread.h>
#define TCOUNT 10
#define WATCH_COUNT 12

int count = 0;
pthread_mutex_t count_mutex = PTHREAD_MUTEX_INITIALIZER;
pthread_cond_t count_threshold_cv = PTHREAD_COND_INITIALIZER;
int thread_ids[3] = {0, 1, 2};

main()
{
  pthread_t threads[3];

  pthread_create((&threads[0], NULL, inc_count, &thread_ids[0]);
  pthread_create((&threads[1], NULL, inc_count, &thread_ids[1]);
  pthread_create((&threads[2], NULL, watch_count, &thread_ids[2]);
  for(i = 0; i < 3; i++)
    pthread_join((&threads[i], NULL);
}
```

```
void watch_count(int *idp)
{
  pthread_mutex_lock(&count_mutex);
  while(count <= WATCH_COUNT) {
    pthread_cond_wait(&count_threshold_cv, &count_mutex);
    printf("watch: Thread %d, Count is %d\n", *idp, count);
  }
  pthread_mutex_unlock(&count_mutex);
}

void inc_count(int *idp)
{
  int i;
  for(i = 0; i < TCOUNT;, i++)
    pthread_mutex_lock(&count_mutex);
    count++;
    printf("inc: Thread %d, count is %d\n", *idp, count);
    if(count == WATCH_COUNT)
      pthread_cond_signal(&count_threshold_cv);
    pthread_mutex_unlock(&count_mutex);
  }
}
```

In this example the reader should note that two threads share identical code and that only one copy of this code is present in memory. The program counter and the stack are of course private property of each individual thread.

A brief resume of the POSIX 1003.1c definitions is given in Annex III. For more details, the reader is referred to the “man pages”.

1.11 Real Time Linux

In October 1996, we learned that a **Real-time Linux** had been developed at the *New Mexico Institute of Technology* and that a *beta-version* was available for testing. Since then Barabanov et. al. have released updated versions of **RTLinux**, and during this College participants can again experiment with it.

It is based on a *different principle* from what we have described so far: it uses the concept of a **virtual machine**. RTLinux embodies a small, **real-time executive**, and standard Linux runs underneath it, as a *low priority task*. The *time critical* parts of the application run directly under `RTLinux` and are scheduled by RTLinux itself. “Classical” Linux is run only when there is *no real-time task ready to run*.

The real-time executive *intercepts the interrupts* and therefore it *can react fast*. Interrupts which have nothing to do with the real-time tasks are passed down to Linux. When Linux *disables interrupts* (with the cli() call), `RTLinux` will stop passing interrupts to Linux. But those interrupts *remain available* to RTLinux for a later time. Linux is used for the lower priority and slower tasks, such as file manipulation. But still all facilities of normal Linux are available.

Communication between a *real-time task* and an *ordinary Linux process* is done via a *special IO interface*, called a **real-time fifo**.

A real-time application should be split into *small and simple parts*, which have *real-time constraints* on the one and larger pieces for more *complex processing* on the other hand.

The real-time component is written as a *dynamically loadable Linux kernel module*. A complete example [1] follows. In this example the real-time part reads periodically data from an external device and puts it into a real-time fifo. The Linux process reads the data from the fifo and can process it. In the example, the data is simply written to *stdout*.

```
#define MODULE
#include <linux/module.h>

/* always needed for real-time task */
#include <linux/rt_sched.h>
#include <linux/rt_fifo.h>
RT_TASK mytask;

/* This is the main program */
void mainloop(int fifodesc)
{
int data;

/* in this loop we obtain data from the */
/* device and put it into fifo number 1 */
while (1) {
    data = get_data();
    rt_fifo_put(fifodesc, (char*)&data, sizeof(data));
    /* give up the CPU until next period */
    rt_task_wait();
    }
}

/* This function is needed in any module */
/* It will be invoked when the module is loaded */
int init_module(void)
{
  #define RTfifoDESC 1
  RTIME now = rt_get_time();

  /* 'rt_task_init' associates a function */
  /* with the RT_TASK structure and sets parameters: */
  /* Priority=4, stack size=3000 bytes, pass 1 to */
  /*'mainloop' as an argument */

  rt_task_init(&mytask, mainloop, RTfifoDESC, 3000, 4);
```

[1] The example is taken from: M. Barabanov and V. Yodaiken, *Introducing Real-Time Linux*, Linux Journal, February 1997, page 19-23.

```
  /* Mark 'mytask' as periodic */
  /* It could be interrupt driven as well */
  /* Period is 25000 time units. It starts */
  /* 1000 time units from now */

  rt_task_make_periodic(&mytask, now+1000, 25000);
  return 0;
}

/* Clean-up routine. It is called when the */
/* module is unloaded */
void clean_up(void)
{
  /* kill the real-time task */
  rt_task_delete(&mytask);
  return;
}
```

The ordinary Linux process executes the following program:

```
#include <rt_fifo.h>
#include <stdio.h>

#define RTfifoDESC 1
#define BUFSIZE 10
int buf[BUFSIZE];

int main()
{
  int i, n;
  /* create fifo number 1, size 1000 bytes */
  rt_fifo_create(1, 1000);
  for (n=0; n<1000; n++) {
    /* read data from fifo and print it */
    rt_fifo_read(1, (char*)buf, BUFSIZE * sizeof(int));
    for (i=0; i<BUFSIZE; i++) {
      printf("%d ", buf[i]);
    }
    printf("\n");
  }
  /* destroy fifo number 1 */
  rt_fifo_destroy(1);
  return 0;
}
```

The latest version of RTLinux can be obtained from http://luz.nmt.edu/řtlinux and besides the executive, it contains kernel patches, documentation, examples and installation tips. Recently, **shared memory** has been added as a means of communicating between RTLinux and standard Linux.

In order to build RTLinux, the Linux kernel must be recompiled. A user can then run either RTLinux or normal Linux at his choice, if lilo.conf is adjusted accordingly.

RTLinux is used at various Institutes around the world. At the Humboldt University in Berlin it is used for data acquisition with an ADC. On a 33 MHz 486 machine, a rate of 3000 samples/s is achieved, using a cheap ADC board connected to the PC via a serial line. At the Universidad Politecnica di Valencia, it is used to develop *earliest deadline first schedules*, including a comprehensive graphical display. Also NASA is using it, but its web sites are not accessible, and we were unable to obtain more information. The New Mexico Institute of Technology itself uses RTLinux in a teaching environment and in the Sunrayce Project (an embedded control system for a solar car?).

The authors of RTLinux say that they could run a repetitive task at a rate of **once every 150 μseconds** on a *133 MHz Pentium*. Tasks can be scheduled within a precision of 10 μseconds.

1.12 RTAI

Paolo Mantegazza and his collaborators at the *Dipartimento di Ingegneria Aerospaziale* of the *Politecnico di Milano* have done a lot of work to improve RTLinux and make it more versatile and user-friendly. The result is a completely new package, **RTAI**, the **Real Time Application Interface**. The package is available from www.aero.polimi.it/projects/rtai. It comes with a large amount of documentation, including explanations of the internals of RTAI, detailed installation procedures, a sort of tutorial and several example programs.

RTAI implements a **hard realtime** system that coexists with Linux. It can run periodic tasks with a frequency exceeding 10 kHz, with a jitter of ± 5 μsec. It can also run in *one-shot* mode.

RTAI has a number of interesting features. It can run on a single processor machine or on a *Symmetric Multi Processor* PC. In the case of SMP, RTAI can be confined to a subset of the processors, or even to a single one. One can also choose to have the realtime interrupts handled by one specific processor, without affecting the interrupts intended for Linux. Pentium or better processors are prefered, but RTAI can run quite reasonably on a 486 machine. To get the most out of it an APIC timer should be available in the processor, besides the usual 8254 timer.

The scheduler functions are also available for the normal Linux processes, which means that you have at your disposal an **uniform API** for all your applications, be they hard, firm or soft realtime. You may use *messages*, *semaphores*, *shared memory* and *time intervals* for communication from *Linux to Linux*, *RTAI to RTAI* and also between *Linux and RTAI*.

When you build RTAI and install it, the realtime application will run in **kernel mode**. You can also make your application run in **user space**, without the need of being the superuser, once the superuser has installed the necessary *kernel modules* for you.

RTAI in fact presents itself in the form of **kernel modules**. Three are required for a basic configuration (comparable to the one provided by RTLinux): the rtai module (rtai), the scheduler (rtl_sched) and the fifo module (rtl_fifo). The application program is added as a fourth module, for instance: rt_process. The latter must be written by the user. A simple example will be shown below. The **fifo** (*First In First Out* buffer) model is the same as for RTLinux: the realtime task or tasks write to a *fifo* and read from another *fifo*. Processes on the Linux side see these *fifo*s as normal *character devices*.

Linux maintains all its features and can thus be used to post-process the acquired data, display the data and archive them, in case we are speaking of a data acquisition system. Likewise, Linux can do also the necessary calculations in the case of a control system. The authors assure that it is possible to run a remote data acquisition system at a rate of one sample every 100 μsec, complete with the network, X11 for displaying results, etc, without Linux falling flat. The authors also suggest that you may use the *interrupt trapping mechanism* on its own, without the scheduler and the rest of the RTAI machinery. This would give you much closer control over the bare hardware, which could be useful for writing a *driver module*.

The following example[1] shows a simple data acquisition application, running in periodic mode at a 10 kHz sampling rate:

```
#define MODULE
#include <linux/module.h>
#include <linux/kernel.h>
#include <linux/version.h>
#include <asm/io.h>
#include <rtai.h>
#include <rtl_sched.h>
#include <math.h>
#include "acquisition_lib.h"

#define STACK_SIZE 2000
#define LOOPS 1000000000

static RT_TASK acquisition_task;

/* This is the function that performs data-acquisition by reading from the
   specific board at a frequency of 10000 Hz */

static void fun_acquis(int t)
{
  unsigned int loops = LOOPS;
  while(loops--){
    read_adc();
    rt_task_wait_period();
  }
}
```

[1] taken from the *"Beginners Guide"* in the RTAI package.

```
int init_module(void)
{
  RTIME now, tick_period;
  int period = 100000;
  rt_set_periodic_mode();   /* The periodic mode is set because
                               we have only one task with a fixed period */
  tick_period = start_rt_timer((int)nano2count(period));
                             /* Conversion of timer period from
                               nanoseconds to internal count units */
  rt_task_init(&acquisition_task,fun_acquis,period,STACK_SIZE, 1, 1, 0);
  now = rt_get_time();
  rt_task_make_periodic(&acquisition_task,now + tick_period,tick_period);
  return 0;
}

void cleanup_module(void)
{
  stop_rt_timer();
  rt_task_delete(&acquisition_task);
}
```

The reader should note that the routine read_adc() has to be written to satisfy the requirements of the specific acquisition board used.

1.13 KURT, the Kansas University Real Time Linux

Very recently, Balaji Shrinavasan of the Information and Telecommunications Technolgy Center (ITTC) of the University of Kansas announced another version of real-time Linux.

It is called **KURT**, for *Kansas University Real Time* Linux.

The author calls it a *firm* real-time system, somewhere between a *hard* and a *soft* real-time system. It is based on a different principle from *RTLinux*.

KURT allows the *explicit scheduling* of real-time **events**, instead of just *processes*. The event scheduling is done by the system.

Once KURT has been installed, Linux has acquired a second mode of operation. The two modes are: *normal-mode* and *real-time mode*. In the first the system behaves as normal Linux, but when the kernel is running in *real-time mode*, it executes only *real-time processes*. All system resources are then dedicated to the real-time tasks. There is a *system call* that toggles between the two modes.

During the *setup phase* the schedule of events to be executed in *real-time mode* is established and the processes that must run in this mode are marked. The kernel is then switched to the *real-time mode*. When all tasks have finished, the kernel is switched back to *normal-mode*.

In order to obtain this behaviour, KURT consists of a **Real-Time Framework** which takes care of scheduling any *real-time event*. When such an event is to be executed, the *real-time framework* calls the **event handler** of the associated

RTMod (Real-Time Module). The *RTMod*s can be very simple; calling them according to a defined schedule is the responsibility of the *real-time framework*. This framework provides the system calls that switch the kernel between the two modes.

The *RTMod*s are *kernel modules*, which are loaded at runtime. An *RTMod* registers itself with the *real-time framework*. It then provides pointers to functions for the *event handler*, *initialization* and *clean-up*.

When a *RTMod* must be invoked is defined in the **Real-Time Schedule**, which is just a **file**. This file can be built beforehand. It can be copied entirely into memory, or it can remain on disk. In the latter case, the timing of events may become distorted by disk access times.

In addition to the *Real-Time Schedule*, *processes* can be run *periodically* in a round-robin fashion.

Events can be scheduled with a high time resolution, when another package has been installed: **UTIME**, for μ*second time*. This package was developed at the same Institute as KURT. If it is not installed then the time resolution is only the usual 10 ms.

To install *UTIME* and/or *KURT*, the Linux kernel must be recompiled.

A more detailed description and *sample programs* are available. The reader should look for them in the directory /usr/local/tarfiles.

The packages together with the necessary Kernel patches can be obtained by anonymous ftp from the WEB page:
`http://www.ittc.ku.edu/kurt/` for *KURT* and
`http://www.ittc.ku.edu/utime/` for *UTIME*.
`http://hegel.ittc.ukans.edu/projects/posix` has extensions to the Linux kernel for better POSIX 1003.1c compatibility.

1.14 Embedded Linux

With portable telephones, handheld and palmtop computers, wireless connections to Internet, automated home appliances and what not, there is a need for **embedded operating systems**. A year or so ago efforts have started to develop these. Several manufacturers have turned to Linux, having recognised that the availability of **open software** bears many advantages, not only to individual developers, hackers and other maniacs, but also for industry. Linux Journal has a regular review on this topic and the September 2000 issue concentrates on the topic. Just to show how active this field is, I cite a few examples from this issue.

- The first example concerns a network of *home infromation appliances*, with a central server and wireless connections to clients scattered through the home. The server obviously runs Linux, but also the clients have an embedded Linux system with a reduced (*small footprint*) X11. Hooking a keyboard, a monitor and a mouse to such a client box, you use it as a PC. Hooking high-fi speakers to it, you listen to music, etc. All the work is done on the server, and data is passed over the network to and from the clients. (http://www.adomo.com).

- *Yopi* is a handheld computer, 12.5 $\times$ 7.5 cm^2 with a color display and a 206 MHz StrongArm CPU with 32 Mbyte DRAM. Another 32 Mbyte of flash ROM stores the Linux system and the core set of applications. The object has no keyboard and looks like a gameboy.

- There is a commercial realtime system based on Linux on the market: *Linux/RT* from *TimeSys Corporation* (email: info@timesys.com). Linux/RT is based on the Carnegie Mellon University *Linux Resource Kernel, Linux/RK*. There is support for *Robust Embedded (RED)* Linux system event logging. Linux/RT also contains RTAI. You run one or the other at any one time. The product may not yet be very mature.

- Aplio bets on voice transmission over Internet (*VoIP*). Their boxes plug into a telephone as simply as an answering machine and the other side of it plugs into an Ethernet port. The boxes run an embedded Linux kernel. It is really based on μClinux, a Linux version for microcontrollers without a memory management unit.

- The firm that produced **LynxOS** (see section 2) changed name. It now calls itself **LynuxWorks**. They released *BlueCat*, a version of Linux tailored for embedded applications. There LynxOS real time operating system will become binary compatible with Linux, so that any executable that runs on Linux can also run on LynxOS.

- Compaq produces *iPaq*, a handheld computer running Linux. The project to develop iPaq originated at Digital Equipment Corporation, which has been bought by Compaq.

- RedHat, in collaboration with Cygnus is working on the development of EL/IX, a Linux-based operating system for embedded applications.

- Lineo is another firm working on Linux for embedded applications.

- To my shame, I don't know where μClinux comes from and what it can do.

1.15 Conclusion

We have tried in this course to give a brief overview of the requirements of a real-time application and we have investigated to what extent Linux can do the job. We have also mentioned the improvements to Linux which have already been made. Among these improvements, `RTLinux` and `KURT` are of the greatest importance for the development of real-time applications. Recent developments in this direction include RTAI and the various projects for *embedded Linux*.

Also the **pthreads** package does contain a major part of the improvements a user would like to see. When a real-time application has been written using *pthreads*, the only essential features the operating system has still to provide are *memory locking* and *real-time scheduling*.

The important thing to remember is that you should analyze your problem very carefully, before deciding that you can (or cannot) use such or such an operating system. Hopefully, this course has shown you the points to consider, and where to search for existing and acceptable solutions.

All depends therefore on your application. If you expect **high data rates** or **high interrupt rates** or if you are otherwise pressed for time constraints, or if you must meet stringent deadlines, then you will need many of the mechanisms described and you should have resort to `RTLinux` or `KURT` or else you may have to accept acquiring a true real-time operating system.

This can be the case in physics experiments, in particular in Particle Physics and in Nuclear Physics.

There will however be situations where you don't need the heavy guns and where the standard Linux system will do the job. To give you an idea: Ulrich Raich runs a real-time application on a 66 MHz 486 machine, concurrently with X11. The machine sustains a rate of 200 external interrupts per second, in addition to the 100 Hz clock interrupts. It obviously all depends on what has to be done as the result of an interrupt.

With prices of PCs and PC-boards going down, there is now a tendency to use a PC-board also for an **embedded system**, where before you would have used a small, dedicated microprocessor. Using a PC-board has the obvious advantage of **portability**: you can develop your application on a large configuration, and then **download** it to the embedded system.

Many people may be just interested in hooking up existing instruments, for instance those which are equipped with an interface to the GPIB bus. This situation arises routinely in chemistry labs, or medical analysis labs, etc. There is good news for those people as well: Packages for controlling instruments with GPIB and Camac exist. The first parts of these were released already in 1995. It has graphical interfaces, uses X11 and is extensible[1]. And **they are free!**. Such packages will certainly deliver the ideal solution for laboratories using standard equipment.

Device drivers for *VME crates and modules* are part of some off-the-shelf Linux distributions (*SuSE 8.1 for instance*).

To end, I wish you a happy time programming your real-time applications! Now that many more tools are available than a few years ago, there is a good chance that this wish comes true.

Enjoy!

[1] You can ftp these packages from koala.chemie.fu-berlin.de.

1.16 Annex I – Annotated bibliography

The last four or five years have seen a flood of books on **Linux**. A number of them are nothing but collections of **HOW-TOs** from the **Linux Documentation Project**. Others are specific for certain **Linux Distributions**, e.g. *Slackware, RedHat, Caldera Desktop, Yggdrasil Plug and Play Linux*. These books generally contain one or more *CD-ROMs*, or the CD-ROM set is sold separately (the *Linux Developers Resource* from *InfoMagic* is an example).

Below is an annotated bibliography of the books I found most useful and which is limited to those publications which are **not** specific to a distribution, or just collections of HOW-TOs. Unfortunately the list has scarcely any item more recent than 1998.

1. Matt Welsh and Lar Kaufman, *Running Linux*, Sebastopol, CA95472, 1995, O'Reilly & Associates, Inc; ISBN 1-56592-100-3
   An excellent book, very complete and very readable. Contains extensive indications on how to obtain and install Linux, followed by chapters on UNIX commands, System Administration, Power Tools (including X11, emacs and LaTeX), Programming, Networking. The annexes contain a wealth of information on documentation, ftp-sites, etc. One of the most readable books on Linux.

2. Marc Ewing, *Running Linux Installation Guide and Companion CD-ROM*, O'Reilly & Associates, Inc.; no apparent ISBN.

3. Matt Welsh, *Linux Installation Guide*, 1995, Pacific Hi-Tech, 3855 South 500 West Suite M, Salt Lake City, Utah 84115,
   email: orders@pht.com; No ISBN found.
   The book is thin (221 pages) and cheap ($ 12.95). It contains a few extra chapters on XFree86, TCP/IP, UUCP, e-mail and usenet.

4. Olaf Kirch, *Linux Network Administrator's Guide*, Sebastopol CA95472, 1995, O'Reilly & Associates, Inc; ISBN 1-56592-087-2
   Another excellent book on Networking for Linux. Covers not only local networks and TCP/IP, but also the use of a serial line to connect to Internet, and other chapters on NFS, Network Information System, UUCP, e-mail and News Readers. Essential reading if you want to use your Linux box on the network.

5. Stefan Strobel and Thomas Uhl, *Linux, unleashing the workstation in your PC*, Berlin, 1994, Springer Verlag; ISBN 3-540-58077-8
   This book is good to whet the appetite of someone who has no idea of what Linux is or what it can do. It has many illustrations, in particular of graphics applications and it mentions many software packages which are not part of the usual Linux distributions, together with indications on how to obtain and install the package.

6. *Linux Bible*, 1994, San Jose, Yggdrasil Computing. No apparent ISBN.
   I know about this book only from the advertisements.

7. Kamram Hussain, Timothy Parker et al., *Linux Unleashed*, 1996, SAMS Publishing, ISBN 0-672-30908-4
   Approx 1100 pages of text, covering Linux and many tools and applications: Editing and typesetting (groff and Tex), Graphical User Interfaces, Linux for programmers (C, C++, Perl, Tcl/Tk, Other languages, Motif, XView, Smalltalk, Mathematics, Database products), System Administration, Setting up an Internet site and Advanced Programming topics. The book contains a CD-ROM with the Slackware distribution.

8. Randolph Bentson, *Inside Linux, a look at Operating System Development*, 1996, Seattle, Specialized system Consultants, Inc; ISBN 0-916151-89-1.
   This book provides some more insight into the internal workings of operating systems, with the emphasis being placed on Linux. It is written in general terms and does not contain code examples.

9. John Purcell (ed.), *Linux MAN, the essential manpages for Linux*, 1995, Chesterfield MI 48047, Linux Systems Lab, ISBN 1-885329-07-5.
   Indispensable for those who cannot stare at a screen for more than 8 hours a day, or who like to sit down in a corner to write their programs with pencil and paper, but want to be sure they use system calls correctly. As the title says, 1200 pages of "man pages" for Linux, from *abort* to *zmore*, and including system calls, library functions, special files, file formats, games, system administration and a kernel reference guide.

10. M. Beck, H. Böhme, M. Dziadzka, U. Kunitz, R. Magnus, D. Verworner, *Linux Kernel Internals*, 1996, Addison Wesley, ISBN 0-201-87741-4.
    There is at least a second edition: ISBN 0-201-33143-8, 1998. For the real sports! A translation of a german book, revealing all the internals of the Linux kernel, including code examples, definitions of structures, tables, etc. The book contains a CD-ROM with Slackware and kernel sources. Indispensable if you want to make modifications to the kernel yourself.

11. Alessandro Rubini, *Linux Device Drivers*, 1998, O'Reilly & Associates, ISBN 1-56592-292-1. This book is a **real must** for anyone wanting to write or modify a device driver for Linux. Before publishing this book, the author had written many articles in the *kernel corner* of *Linux Journal*. The book leads the reader step by step through every corner of a Linux device driver. No secrets are left unveiled.

Having mentioned *Linux Journal*, I should add that you can subscribe via one of the following addresses: e-mail: subs@ssc.com, or on the web: www.linuxjournal.com, Fax: +1-206 297 7515 and by normal mail: SSC, Specialized System Consultants, Inc., PO Box 55549, Seattle, WA 98155-0549, USA. From some thirty pages back in 1994, Linux Journal has grown to around 200

pages monthly. A good fraction of the pages is nowadays occupied by advertisements, but there remain still some 120 or more pages of interesting reading.

Also note that the last few years a number of periodicals on Linux have seen the light in languages other than english.

The following books concern **real-time** and **POSIX.1c**:

i) Bill O. Gallmeister, *POSIX.4: Programming for the Real World*, 1995, O'Reilly & Associates, Inc.; ISBN 1-56592-074-0.
This book gives an in-depth treatment of programming real-time applications, based on the POSIX.4 standard. Several of the examples in the present course were taken from this book. In addition to approximately 250 pages of text, the book contains 200 pages of "man pages" and solutions to exercises.

ii) Bradford Nichols, Dick Buttlar and Jacqueline Proulx Farrell, *Pthreads Programming*, 1996, O'Reilly & Associates, Inc, ISBN 1-56592-115-1.
Probably the best book on Pthreads published so far, concentrating on the POSIX 1003.1c standard and written with a good didactical structure.

iii) Bil Lewis, Daniel J. Berg, *Threads Primer, A Guide to Multithreaded Programming*, 1996, Sunsoft Press (Prentice Hall);
ISBN 0-13-443698-9.
An introduction to threads programming, mainly based on the Solaris implementation of threads, but containing comparisons to POSIX threads and a full definition of the *Applications Programmer's Interface* to **POSIX.4a pthreads**.

iv) S. Kleiman, Devang Shah, B. Smaalders, *Programming with Threads*, 1996, Sunsoft Press (Prentice Hall; ISBN 0-13-172389-8. This book contains a more in-depth treatment of threads programming than the previous title. It is also more pthreads-oriented.

v) Andrew S. Tanenbaum, *Modern Operating Systems*, 1992, Prentice Hall; ISBN 0-13-595752-4.
This excellent book is not specifically tuned to real-time, but it provides a comprehensive introduction to the features of modern operating systems and their implementation. An older edition of the book contained a complete listing of the **minix** operating system. The reader may appreciate that Linux was born when Linus Torvalds set out to improve minix . . .

vi *Last minute addition:* O'Reilly is expected to issue a *community written* book on Linux realtime. The editor is *Phil Daly* of *realtimelinux.org*. The announcement says: *"The volume will be the definite guide to the installation and use of real time Linux and will feature a bootable CD-ROM to help "get you going" with this exciting technological development. The text will be made available under an Open Content License agreement with content under constant review."*

Note that there are many more books available, in particular from O'Reilly, which may be of relevance to topics treated in the present course. Finally, there is a paper on Real Time Linux:
M. Barabanov and V. Yodaiken, *Introducing Real-Time Linux*, Linux Journal, February 1997, pages 19-23.

For RTAI, there is a large amount of documentation available from:
http://www.aero.polimi.it/projects/rtai/.

To conclude, some mailing lists which may be useful. To subscribe to any of the lists, send an email to: **missdomo@realtimelinux.org** with **subscribe 'listname'** in the **body** of the email.

- **realtime** — realtime@realtimelinux.org — general discussion.
- **api** — api@realtimelinux.org — API discussion.
- **documentation** — documentation@realtimelinux.org
- **drivers** — drivers@realtimelinux.org — discussion of drivers for use with Realtime Linux.
- **kernel** — kernel@realtimelinux.org — Linux kernel modifications for use with Realtime Linux.
- **networking** — networking@realtimelinux.org — Realtime networks.
- **ports** — ports@realtimelinux.org — Porting of RTL/RTAI to other platforms.
- **testing** — testing@realtimelinux.org — discussion on testing

1.17 Annex II – CD-ROM sets

All the well-known Linux distributions (*Caldera, Corel, Debian, RedHat, Slackware, SuSE* and I will certainly miss out a few...) now come on two or more CD-ROMs. Beside the base system they contain in general a wealth of additional, optional packages and a lot of documentation.

1.18 Annex III — Resume of POSIX 1003.1c definitions

POSIX.1c/D10 Summary

Disclaimer

SunSoft, Inc.
2550 Garcia Avenue
Mountain View, California 94043

Introduction

All source that uses POSIX.1c threads must include the header file.

```
#include <pthread.h>
```

In addition, Solaris requires the pre-processor symbol **_REENTRANT** to be defined in the source code before any C source (including header files).

```
#define _REENTRANT
```

The POSIX.1c thread library should be the last library specified on the cc(1) command line.

```
voyager% cc -D_REENTRANT ...  -lpthread
```

Name Space

Each POSIX.1c type is of the form:

```
pthread[_object]_t
```

Each POSIX.1c function has the form

```
pthread[_object]_operation[_np|_NP]
```

where `object` is a type (not required if object is a thread), `operation` is a type-specific operation and np (or `NP`) is used to identify non-portable, implementation specific functions.

All POSIX.1c functions (except for `pthread_exit`, `pthread_getspecific` and `pthread_self`) return zero (0) for success or an `errno` value if the operation fails.
There are eight(8) POSIX.1c types, see table 1.1:

Table 1.1: POSIX.1c types

Type	**Description**
pthread_attr_t	Thread attribute
pthread_mutexattr_t	Mutual Exclusion Lock attribute
pthread_condattr_t	Condition variable attribute
pthread_mutex_t	Mutual Exclusion Lock (mutex)
pthread_cond_t	Condition variable (cv)
pthread_t	Thread ID
pthread_once_t	Once-only execution
pthread_key_t	Thread Specific Data (TSD) key

Feature Test Macros

POSIX.1c consists of a base (or common) component and a number of implementation optional components The base is the set of required operations to be supplied by every implementation. The preprocessor symbol (**_POSIX_THREADS**) can be used to test for the presence of the POSIX.1c base. Additionally, the standards document describes a set of six (6) optional components. A pre-processor symbol can be used to test for the presence of each All of the symbols appear in the table 1.2.

Table 1.2: POSIX.1c Feature Test Macro

Feature Test Macro	**Description**
_POSIX_THREADS	base threads
_POSIX_THREAD_ATTR_STACKADDR	stack address attribute
_POSIX_THREAD_ATTR_STACKSIZE	stack size attribute
_POSIX_THREAD_PRIORITY_SCHEDULING	thread priority scheduling
_POSIX_THREAD_PRIO_INHERIT	mutex priority inheritance
_POSIX_THREAD_PRIO_PROTECT	mutex priority ceiling
_POSIX_THREAD_PROCESS_SHARED	inter-process synchronization

Macro Dependency

If **_POSIX_THREAD_PRIO_INHERIT** is defined
then **_POSIX_THREAD_PRIORITY_SCHEDULING** is defined.

If **_POSIX_THREAD_PRIO_PROTECT** is defined
then **_POSIX_THREAD_PRIORITY_SCHEDULING** is defined.

If **_POSIX_THREAD_PRIORITY_SCHEDULING** is defined
then **_POSIX_THREADS** is defined.
If **_POSIX_THREADS** is defined
then **_POSIX_THREAD_SAFE_FUNCTIONS** is defined.

POSIX.1c API

In the following sections, function arguments that are of the form:

```
type name = NULL
```

indicate that a value of **NULL** may safely be used for name.

```
int pthread_atfork( void (*prepare)(void) = NULL,
      void (*parent)(void) = NULL,
      void (*child)(void) = NULL );
```

Register functions to be called during fork execution.
errors **ENOMEM**
notes `prepare` functions are called in reverse order of registration.
`parent` and `child` functions are called in order of registration.

Thread Attributes

All thread attributes are set in an attribute object by a function of the form:

```
int pthread_attr_setname( pthread_attr_t *attr, Type t );
```

All thread attributes are retrieved from an attribute object by a function of the form:

```
int pthread_attr_getname( const pthread_attr_t *attr, Type *t );
```

Where `name` and `Type` are from the table 1.3.

`int pthread_attr_init( pthread_attr_t *attr );` Initialize a thread attribute object.
errors **ENOMEM**

`int pthread_attr_destroy( pthread_attr_t *attr );`
Destroy a thread attribute `object` .
errors none

Table 1.3: Thread Attributes

Name and Type	Feature Test Macro	Value(s)
int inheritsched	**_POSIX_THREAD_PRIORITY_SCHEDULING**	**PTHREAD_INHERIT_SCHED, PTHREAD_EXPLICIT_SCHED**
int schedpolicy	**_POSIX_THREAD_PRIORITY_SCHEDULING**	**SCHED_FIFO, SCHED_RR , SCHED_OTHER**
struct sched_param schedparam	**_POSIX_THREADS**	POSIX.1b, Section 13
int contentionscope	**_POSIX_THREAD_PRIORITY_SCHEDULING**	**PTHREAD_SCOPE_SYSTEM, PTHREAD_SCOPE_PROCESS**
size_t stacksize	**_POSIX_THREAD_ATTR_STACKSIZE**	>= **PTHREAD_STACK_MIN**
void *stackaddr	**_POSIX_THREAD_ATTR_STACKADDR**	void *stack
int detachstate	**_POSIX_THREADS**	**PTHREAD_CREATE_DETACHED, PTHREAD_CREATE_JOINABLE**

Thread Management

```
int pthread_create( pthread_t *thread,
       const pthread_attr_t *attr =  NULL,
       void *(*entry)(void *), void *arg );
```

Create a new thread of execution.
errors **EAGAIN, EINVAL**
note **Maximum number of PTHREAD_THREADS_MAX threads per process.**

```
int pthread_detach( pthread_t thread );
```

Set the detachstate of the specified thread to **PTHREAD_CREATE_DETACHED.**
errors **EINVAL, ESRCH**

```
pthread_t pthread_self( void );
```

Return the thread ID of the calling thread.
errors none

```
int pthread_equal( pthread_t t1, pthread_t t2 );
```

Compare two thread IDs for equality.
errors none

```
void pthread_exit( void *status = NULL );
```

Terminate the calling thread.
errors none

```
int pthread_join( pthread_t thread, void **status = NULL );
```

Synchronize with the termination of a thread.
errors **EINVAL, ESRCH, EDEADLK**
note This function is a cancellation point.

#include <sched.h>

```
int pthread_getschedparam( pthread_t thread, int *policy, struct sched_param
   *param );
```

Get the scheduling policy and parameters of the specified thread.
control **_POSIX_THREAD_PRIORITY_SCHEDULING**
errors **ENOSYS, ESRCH**

#include <sched.h>

```
int pthread_setschedparam( pthread_t thread, int policy,


       const struct sched_param *param );
```

Set the scheduling policy and parameters of the specified thread.
control **_POSIX_THREAD_PRIORITY_SCHEDULING**
errors **ENOSYS, EINVAL, ENOTSUP, EPERM, ESRCH**
policy { **SCHED_RR, SCHED_FIFO, SCHED_OTHER** }

Mutex Attributes

All mutex attributes are set in a mutex attribute object by a function of the form:

```
int pthread_mutexattr_setname( pthread_attr_t *attr, Type t );
```

All mutex attributes are retrieved from a mutex attribute object by a function of the form:

```
int pthread_mutexattr_getname( const pthread_attr_t *attr, Type *t
);
```

Where name and Type are from the table 1.4

Table 1.4: Mutex Attributes

Name and Type	**Feature Test Macro**	Value(s)
`int protocol`	**_POSIX_THREAD_PRIO_INHERIT**	**PTHREAD_PRIO_NONE,**
	_POSIX_THREAD_PRIO_PROTECT	**PTHREAD_PRIO_PROTECT,**
		PTHREAD_PRIO_INHERIT
`int pshared`	**_POSIX_THREAD_PROCESS_SHARED**	**PTHREAD_PROCESS_SHARED,**
		PTHREAD_PROCESS_PRIVATE
`int prioceiling`	**_POSIX_THREAD_PRIO_PROTECT**	POSIX.1b, Section 13

```
int pthread_mutexattr_init( pthread_mutexattr_t *attr );
```

Initialize a mutex attribute object.
errors **ENOMEM**

```
int pthread_mutexattr_destroy( pthread_mutexattr_t *attr );
```

Destroy a mutex attribute object.
errors **EINVAL**

Mutex Usage

```
int pthread_mutex_init( pthread_mutex_t *mutex,
   const pthread_mutexattr_t *attr = NULL );
```

```
pthread_mutex_t mutex = PTHREAD_MUTEX_INITIALIZER;
```

Initialize a mutex.
errors **EAGAIN, ENOMEM, EPERM, EBUSY, EINVAL**

```
int pthread_mutex_destroy( pthread_mutex_t *mutex );
```

Destroy a mutex.
errors **EBUSY, EINVAL**

```
int pthread_mutex_getprioceiling( const pthread_mutex_t *mutex,
   int *prioceiling );
```

Get the prioceiling value of the specified mutex.
control **_POSIX_THREAD_PRIO_PROTECT**
errors **ENOSYS, EINVAL, EPERM**

```
int pthread_mutex_setprioceiling( pthread_mutex_t *mutex, int prioceiling,
       int *old_ceiling );
```

Set the prioceiling value and return the old prioceiling value in the specified mutex.
control **_POSIX_THREAD_PRIO_PROTECT**
errors **ENOSYS, EINVAL, EPERM**

```
int pthread_mutex_lock( pthread_mutex_t *mutex );
```

Acquire the indicated mutex.
errors **EINVAL, EDEADLK**

```
int pthread_mutex_trylock( pthread_mutex_t *mutex );
```

Attempt to acquire the indicated mutex.
errors **EINVAL, EBUSY, EINVAL**

```
int pthread_mutex_unlock( pthread_mutex_t *mutex );
```

Release the (previously acquired) mutex.
errors **EINVAL, EPERM**

Once-only Execution

```
pthread_once_t once = PTHREAD_ONCE_INIT;
```

Initialize a once control variable.

```
int pthread_once( pthread_once_t *once_control, void (*init_routine)(void)
   );
```

Execute init_routine once.
errors none specified

Condition Variable Attributes

All condition variable attributes are set in a condition variable attribute object by a function of the form:

```
int pthread_condattr_setname( pthread_condattr_t *attr, Type t );
```

All condition variable attributes are retreived from a condition variable attribute object by a function of the form:

```
int pthread_condattr_getname( const pthread_condattr_t *attr, Type
*t );
```

Where name and Type are from the table 1.5

Table 1.5: Condition Variable Attributes

Name and Type	Feature Test Macro	Value(s)
int pshared	**_POSIX_THREAD_PROCESS_SHARED**	**PTHREAD_PROCESS_SHARED, PTHREAD_PROCESS_PRIVATE**

`int pthread_condattr_init( pthread_condattr_t *attr );`
Initialize a condition variable attribute object.
errors **ENOMEM**

`int pthread_condattr_destroy( pthread_condattr_t *attr );`
Destroy a condition variable attribute object.
errors **EINVAL**

Condition Variable Usage

```
int pthread_cond_init( pthread_cond_t *cond,

       const pthread_condattr_t *attr = NULL );

pthread_cond_t cond = PTHREAD_COND_INITIALIZER;
```

Initialize a condition variable.
errors **EAGAIN, ENOMEM, EBUSY, EINVAL**

`int pthread_cond_destroy( pthread_cond_t *cond );`
Destroy a condition variable.
errors **EBUSY, EINVAL**

`int pthread_cond_signal( pthread_cond_t *cond );`
Unblock at least one thread currently blocked in the specified condition variable.
errors **EINVAL**

`int pthread_cond_broadcast( pthread_cond_t *cond );`
Unblock all threads currently blocked on the specified condition variable.
errors **EINVAL**

`int pthread_cond_wait( pthread_cond_t *cond, pthread_mutex_t *mutex );`
Block on the specified condition variable.
errors **EINVAL**
note This function is a cancellation point.

```
int pthread_cond_timedwait( pthread_cond_t *cond, pthread_mutex_t *mutex,
       const struct timespec *abstime );
```

Block on the specified condition variable not longer than the specified absolute time.
errors **ETIMEDOUT, EINVAL**
note This function is a cancellation point.

Thread Specific Data

```
int pthread_key_create( pthread_key_t *key, void (*destructor)(void *) = NULL
    );
```

Create a thread-specific data key.
errors **EAGAIN, ENOMEM**
note system limit of PTHREAD_KEYS_MAX per process.
system limit of PTHREAD_DESTRUCTOR_ITERATIONS calls to destructor per thread exit.

```
int pthread_key_delete( pthread_key_t key );
```

Destroy a thread-specific data key.
errors **EINVAL**

```
void *pthread_getspecific( pthread_key_t key );
```

Return the value bound to the given key for the calling thread.
errors none

```
int pthread_setspecific( pthread_key_t key, const void *value );
```

Set the value for the given key in the calling thread.
errors **ENOMEM, EINVAL**

Signal Management

```
#include <signal.h>

int pthread_sigmask( int how, const sigset_t *newmask = NULL,
    sigset_t *oldmask = NULL );
```

Examine or change calling threads signal mask.
errors **EINVAL**
how { **SIG_BLOCK, SIG_UNBLOCK, SIG_SETMASK** }

```
#include <signal.h>

int pthread_kill( pthread_t thread, int signo );
```

Deliver signal to indicated thread.
errors **ESRCH, EINVAL**

```
#include <signal.h>

int sigwait( const sigset_t *set, int *sig );
```

Synchronously accept a signal.
errors **EINVAL, EINTR**
note This function is a cancellation point.

Cancellation

`int pthread_setcancelstate( int state, int *oldstate );`
Set the cancellation state for the calling thread.
errors **EINVAL**
`state` { **PTHREAD_CANCEL_ENABLE, PTHREAD_CANCEL_DISABLE** }

`int pthread_setcanceltype( int type, int *oldtype );`
Set the cancellation type for the calling thread.
errors **EINVAL**
`type` {**PTHREAD_CANCEL_DEFERRED, PTHREAD_CANCEL_ASYNCHRONOUS** }

`int pthread_cancel( pthread_t thread );`
Cancel the specified thread.
errors **ESRCH**
note threads that have been cancelled terminate with a status of **PTHREAD_CANCELED.**

`void pthread_testcancel( void );`
Introduce a cancellation point.
errors none
note This function is a cancellation point.

`void pthread_cleanup_pop( int execute );`
Pop the top item from the cancellation stack and optionally execute it.
errors none specified
note push and pop operations must appear at the same lexical level.
`execute` { `1, 0` }

`void pthread_cleanup_push( void (*routine)(void *), void *arg );`
Push an item onto the cancellation stack.
errors none specified