An Example Design of a Non-preemptive Real-Time Kernel

Version 1.0

This is a preliminary draft of the document. It may be missing information and is incomplete. Please report any errors or omissions.

P. Dasiewicz © 2007

Introduction

This version of the document is not meant to function as a standalone all inclusive beginner's guide to the design of an embedded real-time executive (the kernel). It is to accompany the RTX lecture slides and it is assumed that at least the first few chapters of the course text have been covered in lectures. A possible design for a basic non-preemptive message passing RTX is initially presented and later extended to account for process preemption. Possible pseudocode is provided as implementation examples as appropriate and many course project related issues are discussed but without detailed solutions – which would be expected to be solved during the the course project. Later sections overview the UNIX run-time environment with detailed examples of performing process switching, forking helper processes and using UNIX signals needed for the course project. Some sections illustrate the application of the concepts to a specific processor, the MCF 5307.

Table of Contents

1 RTX Functional Overview

This document is a reference text to the RTX lecture slides. First, we overview the basic desired functionality and any limitations of our basic non-preemptive real time executive. In later sections, the basic design is extended to include preemption. The material is presented from a generic point of view, where concepts are valid for any embedded real-time system. Also, some examples are given for a specific MCF5307 CPU application to indicate how some design decisions are driven by the underlying hardware architecture. Finally, some additional constraints are discussed which are driven by the specific MTE 241 project specifications.

1.1 Process Management

At the very least we need to be able to create and terminate user processes. Since we are considering an embedded application, the ability to terminate a user process is not needed since the assumption is that all processes exist for the lifetime of the system once they are created. A second requirement is to perform a complete system initialization at system start (power-up). This would include initialization of all hardware, creation of all processes, initializing the required stack and heap spaces.

1.2 Processor Scheduling

We need to select a process that is ready to execute and change it to executing. A criteria must be defined to chose among many ready processes.

1.3 Interprocess Communication and Synchronization

Some processes need to communicate and synchronize their actions with each other to perform some overall functionality. The communication and synchronization mechanisms must be defined.

1.4 Storage Management

A mechanism to allocate and deallocate memory requests must be provided. The execution times for storage management functionality must be “predictable” and short. This mechanism will be customized to the requirements of the intended application to reduce the overhead costs.

1.5 Interrupt Handling Framework

Capture interrupts and, if required, start (or make ready) a user process that was blocked, waiting for the event to occur.

1.6 Timing Services

Since real-time is the application, the RTX must provide services which measure time or occur at specific points in time. There may be a need for either (or both) absolute and relative timing requests.

1.7 Device Driver Interfaces

There is need to provide standard I/O and specialized interrupt driven device handlers.

2 Simple RTOS: Requirements

To reduce complexity, the following basic requirements will drive the design. Initially, the assumption is the RTX is to be non-preemptive. This constraint is initially imposed to reduce the initial complexity of the design and will be removed near the end. Since the intended application is embedded real-time, the requirement of process creation after initialization is not needed. Also, no process terminates during the lifetime of the RTX. These two requirements greatly simplify the design. In a general purpose OS, these requirements would not be appropriate and never made.

We assume that the priority (process urgency) is sufficient to use as a scheduling criteria. Further, process priority is static and does not vary dynamically.

The IPC mechanism is to be communication via message envelopes. Processes will send and receive message envelopes (shared blocks of memory) to communicate and synchronize their actions. Sending a message will be asynchronous while receiving a message will be synchronous.

A simple form of memory management will be used. Fixed size blocks of memory will be allocated/deallocated by processes and used as message envelopes for IPC. This avoids the extra complexity and overhead of variable sized memory allocation/deallocation.

There is the need for only a basic form of timing services: relative time.

Some additional assumptions which greatly simplify the RTX complexity include the following:

- All processes are known and created at initialization time. This removes any requirement to dynamically create processes.
- All processes are “friendly, co-operating and non malicious”. This produces considerable complexity reduction since most 'security' related issues can be ignored. The RTX trusts each process to do the right thing and user processes will never do anything that could harm the RTX or each other. No general purpose OS would ever make such an assumption. For example, there is no need to track memory allocated to a process since the process will always return it. Since only a non-preemptive design is initially considered, this would also imply that a process would voluntarily yield the processor at designated points in its code in order not to monopolize the processor.
- For a process to send a message to another process, the sender must know the process id of the receiving process. The assumption is that all processes know the process id of all other processes and hence no RTX primitives need be provided for a process to determine its own process id or that of any other process.

3 The User Process / Kernel Interface

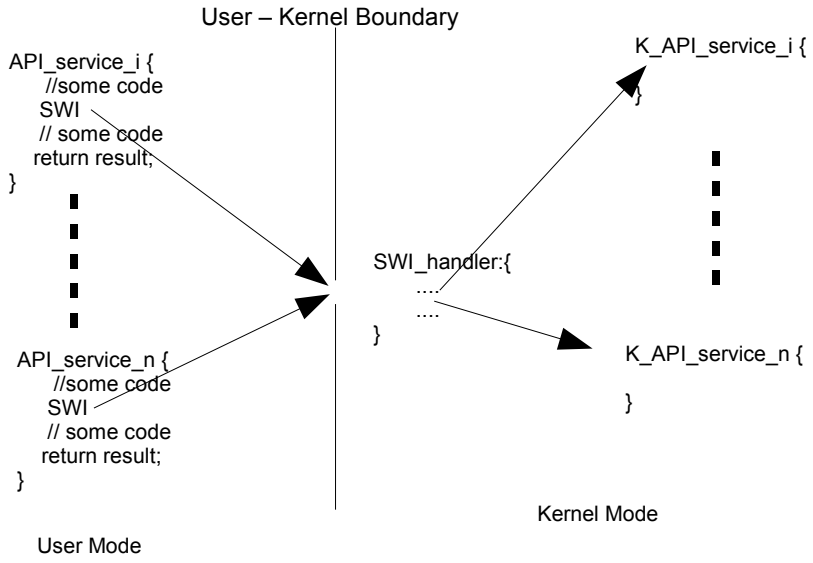
A user process executes in a restrictive mode (internal kernel data structures are not accessible or visible, usually with only a subset of the CPU instructions available, a restricted address space and often with a separate user stack,...). This execution mode is usually referred to as *user mode*. The kernel executes in the CPU's *supervisor mode* (also referred to as *kernel mode*) and can execute all the CPU's instruction set and has access to the complete address space with no restriction on its operation. When a user process is scheduled (by the kernel) to execute, then the kernel ensures that the user process only executes in *user mode*.

The kernel provides a set primitives (services) to user processes. A user process is provided a user visible API. The API defines all services (and their signature) that the kernel provides. As seen in later sections, the actual service (functionality) of each API service must be executed in kernel mode. Thus, the initial problem is to determine the actual user process / kernel interface. What happens when a user process invokes an API service? Since the API defines only user visible services, the actual function invoked is still in user mode and cannot by

itself provide the required service since the service can only be run in supervisor mode , that is, kernel mode.

Fundamentally, in response to the user process' call to a visible API service, the service performs the following operations: i) repackages the passed parameters, ii) enters the kernel by a software interrupt, iii) extracts any return value from kernel, and iv) performs a return to the user process. Usually, the kernel has a single entry point which is entered by issuing a software interrupt (SWI) instruction (also known as a TRAP instruction). This causes an exception handling sequence to be performed by the CPU and control is transferred to the SWI handler provided by the kernel. The SWI handler (executing now in kernel mode) will determine which kernel service was requested and perform the service in kernel mode. At the completion of the service, control is transferred back to the user visible API service and it resumes execution in its original user mode. See Appendix A for a detailed example.

The following figure illustrates the basic user process API to kernel interface. The kernel version of the user visible API is preceded by a "K_" to indicate that this is the actual kernel implemented service. A short example will illustrate the mechanism.



The following generic example is based on the following assumptions:

1. No special knowledge of the C compiler, its run-time stack frames or assembler is assumed.
2. No particular processor is assumed except that it can execute SWI's; and that it has defined data registers named D0, D1, ...
3. On a SWI, the D0 register will hold a constant denoting the required kernel primitive. And on return from SWI, the register D0 will hold any returned value.
4. Assume that the function ASM("*processor instruction*") is known to the compiler and will cause it to place the processor instruction into the produced assembly language at the given point.

Consider the following user visible API function for a send message primitive (discussed later):

```
int send_message( int dest_pid, char * msg )
```


This API is a request to send a message to a process with process id *dest_pid* and it returns an integer result.

A generic C implementation of send_message at the user level could be:

```
int send_message( int dest_pid, char * msg)
{
       int result_value;

       ASM( "PUSH D2");          // save on stack some registers
       ASM( "PUSH D1");
       ASM( "PUSH D0");
       ASM( "MOVE dest_pid, D1");
       ASM( "MOVE msg, D2");
       ASM( "MOVE #K_SEND_MSG, D0");  //the SWI handler will look at D0
       ASM( "SWI" );                  //enter kernel here
       ASM( "MOVE D0, result_value"); // save the returned value
       ASM( "POP D0");                // restore registers from stack
       ASM( "POP D1");
       ASM( "POP D2");
       return result_value ;
}
```

The generic example illustrates the basic outline used for all user visible API functions. This can be optimized once the C compiler and its run-time stack frames are known. Other user visible API functions would be implemented as above with the basic difference being the number of parameters in the API function itself and the return type of the result.

We can now determine a generic outline for the kernel's SWI handler. All user API's perform a SWI to the same handler. The SWI handler determines the nature of the request by examining the D0 register which would have a different value for each user API service. The following illustrates a basic generic SWI handler:

```
SWI_handler:
{
       int int_var1, int_var2, ret_value;
       char * ptr_1;
       char * ptr_2;

       // some initial code to save appropriate registers, etc....
       // all done in assembly language (also disabling interrupts?)
       ASM("MOVE D1, int_var1"); //copy the dest_pid from user
       ASM("MOVE D2, ptr_1");    // the ptr to the message from user
       ASM("MOVE D0, int_var2");  // copy the service identifier
       switch( int_var2 ) {
              case  K_SEND_MSG : // the send message was invoked!
                                 ret_value = K_send_message( int_var1, ptr_1);
                                 break;
              case other_selections :  // other kernel services as needed
       };  //end off the switch statement
       // Now restore any saved registers and prepare for return to user.
       // Add any assembly language here to restore registers saved earlier.
       // put the return value into register D0 for the user API
       ASM("MOVE ret_value, D0");
       ASM( " RTE "); // the rte will now restore the state of the processor
                      // back to the statement following the original SWI
}
```

The entry to SWI handler 'appears' to be customized to the send_message API, but in general, it will work for all user API calls which have at most 2 formal parameters. The actual type conversion from the entry code to the actual parameters required by each kernel API function will be performed individually (perhaps using 'casts') by the respective 'case' alternative of the 'switch' statement. On return to the user API function (after the SWI), register D0 will have the appropriate return value and this will be returned to the user process (as seen previously).

4 Requirement for Atomicity

There is a requirement for each RTX primitive to be atomic, that is all instructions of any primitive must execute indivisibly with no interruption. Once a kernel primitive starts, it runs to completion with no interruption. This is required since primitives access and modify private kernel variables and data structures. It would not be permissible for an interrupt handler to access/modify the same kernel data structures concurrently with a primitive's execution. Thus, it must be ensured that all (or least any potentially interfering) interrupts be disabled during the execution of the primitive's code.

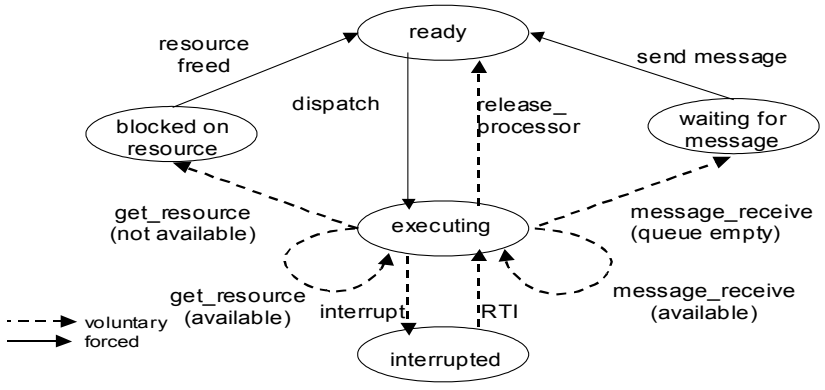
In reality, all the primitive functionality will be performed in "kernel mode" (with interrupts disabled) by kernel versions of the user visible API functions, after the user API 'traps' into the kernel. On entry, the software trap (SWI) handler ensures that interrupts are disabled while the kernel primitive is executing.

Since the primitives (services) run in kernel mode (after a SWI from the user visible API), it is simple to disable all (or some potentially interfering) interrupts by the SWI handler. Recall that in kernel mode, it is now possible to execute privileged CPU instructions to disable/enable interrupts. After the service completes, then the original interrupt masks (used in user mode) can be restored prior to returning to the user process.

All kernel primitive implementations, in following sections, assume that execution is in kernel mode and that atomicity is guaranteed.

5 Process State Diagram

Since a process can be in various states during its execution, the process state diagram is required. The following is a typical state diagram that can be used in this example design. While in the executing state, a process can voluntarily yield the processor (via the 'release_processor' primitive) and return back to the ready state allowing another process to be selected for execution. If a process requests a message_receive and no message is available, then the process blocks and is placed into the waiting_for_message state until another process sends it a message; after which it returns to the ready state. If a process requests a resource (get_resource primitive) and no resource is available, then the process blocks and is placed in the blocked_on_resource state until a resource becomes available. The dispatch action of the kernel selects a ready process for execution and changes its state



to executing. A process moves from executing to interrupted state during interrupt handling and returns to executing state after the interrupt handler completes. Since our initial design is non-preemptive, a process always returns to executing from the interrupted state. For a preemptive system, some changes would be required to the state diagram.

6 Kernel Private Data Structures/Functions

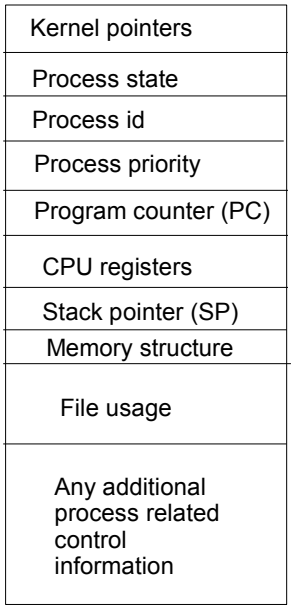
Some of the private kernel variables, data structures and functions are examined. These structures are not visible nor accessible by a user process.

6.1 The Process Control Block (PCB)

In order to manipulate a process (i.e. make it ready, block it, put a process on a queue, etc), the kernel needs an internal representation of a process. A data structure that contains process related information is required. It is this data structure that is placed on linked lists etc. not the physical process itself. Typically, a process control block (PCB) or a process object is used to represent a process. The PCB is not accessible to a user process since it is a private data structure maintained by the kernel. The following figure illustrates a typical PCB data structure which can be used. Depending on the application, the PCB can contain many different fields, some of which are discussed.

Each process (user process, system process or i_process) requires a PCB which is constructed during process creation and remains in existence until the process terminates. The PCB contains several kernel pointer fields allowing the PCB to be placed on one or more linked lists simultaneously. For example, each PCB is permanently placed on the system process list which is a linked list of all processes in the system. In addition, the PCB may also be placed on various blocked queues and ready queues.

The "process state" field is used to indicate the current state of the process (i.e. ready, blocked-on-receive, executing, suspended, ...).This field is maintained by the kernel during the lifetime of the process. Kernel primitives examine and modify this field, e.g. when a process is made ready or is blocked.



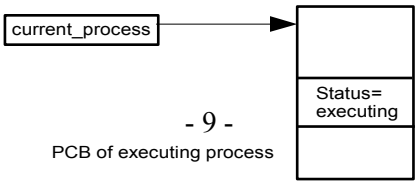
The "process id" field holds a unique identifier for each process. This is usually a unique integer number assigned to each process and not reused again. The "process priority" field contains the priority or urgency of a process and is used to schedule processes and to determine if process preemption is required. This is usually an integer in some restricted subrange.

The next three fields "PC, CPU registers, SP" are used to maintain the process context when the process is not executing. The process context (CPU registers) can be saved here when the process is not running and can be used to restore a process to a running state. These fields are only valid when a process is not running and its context has been saved.

The memory structure field can be used to maintain information about memory allocated to the process (for example in a VM environment). The "file usage" field can be used to record which files have been opened by the process and the state of each file. Additional fields can be used to maintain various statistics required by the kernel (total execution time of process, waiting time, number of page faults, etc...).

6.2 The *current_process* Variable

The kernel must know the identity of the currently executing process at all times. For example, when a process invokes a kernel primitive, the primitive must be able to deduce the identity of the requesting process; also an interrupt handler might need the identity of the



process that was executing at the time of the interrupt. The kernel maintains a variable called `current_process` which points to the PCB of the currently executing process (see adjacent figure). This variable is updated each time a different process is chosen for execution.

6.3 The *process_switch*() Function

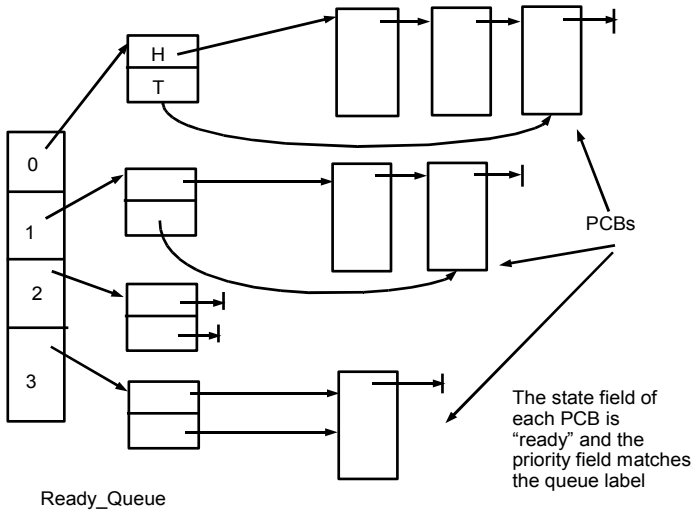
This is a frequently needed procedure to i) remove the current process from the CPU, ii) select the next process to execute, and iii) transfer control to the selected process (restart it). For example, an RTX primitive decides to block the requesting process and to select another process to run. Alternatively, in a preemptive system, an interrupt handler makes ready a process of higher priority than the currently executing process, and decides to switch to the newly ready'ed process.

For example, before calling `process_switch`(), a primitive would have placed the blocked process on a blocked queue (if appropriate) and made any needed update to the state field of the current process's PCB. No return to the call is made until the original process is restarted again in the future. In this example, process switch would call the scheduler to select a new process to run. Then, a call is made to the function `context_switch(new_pcb_ptr)` passing to it a pointer to the PCB of the newly selected process to execute. Context_switch performs the following: i) saves the context of the currently executing process (if not done already), ii) sets the variable `current_process` to refer to `new_pcb_ptr` , iii) sets the state of the new `current_process` to executing, iv) restores context of `current_process`, and v) restarts the execution of `current_process`.

After a primitive invokes `process_switch`, the invoking process will eventually regain execution again and will restart on the statement following the call to `process_switch`. See a later section for a more detailed look at this functionality.

6.4 Process Scheduling

The requirement specified a fixed priority based approach where each process is assigned a priority (a process urgency). The highest priority ready process will be chosen to execute. Processes with equal priority are treated as first-come-first-served (FCFS). This priority based scheduling is applicable to both pre-emptive and non pre-emptive applications. As with any fixed priority scheduling, there is always a danger of starvation for the low priority processes if there are too many higher priority processes ready to run. Other scheduling algorithms are used in more complex real-time systems. The following figure shows a 4 level priority ready queue with the highest priority level being '0'.



6.4.1 The *rpq_enqueue/dequeue* Functions

There is also need of enqueuing and dequeuing PCBs to/from the ready process queue (for example, the previous figure). `rpq_enqueue(pcb_ptr)` enqueues the specified PCB onto the appropriate ready queue by obtaining the process's priority from the priority field in the PCB and inserting the PCB to the end of the respective queue.

`rpq_dequeue()` is a function that returns a pointer to the PCB of the highest priority ready process. The PCB is also removed from the respective queue. The function scans the four queues (queue 0 is the highest priority) and determines the highest priority ready process that is to be run next and removes its PCB from the respective queue.

6.4.2 The Null Process

What happens if the scheduler cannot find a ready process? This could occur if all the ready queues are empty. We cannot just wait inside the kernel for some process to become ready. Since all processes appear to be blocked, then either the system is in deadlock or one or more processes are waiting for some I/O device to complete and the appropriate interrupt handler might make a blocked process ready and save the day. This condition must be accounted for in the design. There are two fundamental alternatives in this case.

First, we could intelligently loop inside the kernel waiting for an interrupt to make a process ready. For example consider the following pseudo code:

```
while (ready_Q is empty ) do
    EI              //enable interrupts
    DI              // disable interrupts
endLoop
```

The loop enables interrupts for a brief time period such that if there is a pending interrupt, then the processor will recognize the interrupt, perhaps resulting in an interrupt handler making a blocked process ready. With this approach, the interrupt handlers must be aware that the kernel is in this loop when the handler runs otherwise the handler might force a process switch with some unforeseen consequences. This approach is not recommended for the project although there are some beneficial results in this approach from the performance point of view.

An alternative approach might be to include a special *null process* that is always ready to run and never blocks

on any kernel primitive. The null process's priority is set to the lowest priority level such that it should not be selected by the scheduler unless there are no other ready processes. The null process executes in user mode. When the ready queue is "empty" of user processes, then the scheduler would pick the null process to execute next.

A simple implementation of the null process (for a non preemptive system) could be as:

```
null_process( )
{
       while ( true )  {
            release_processor( );
       }
}      //end of null_process
```

The null process consists of an infinite loop with a call to the `release_processor( )` API service. The `release_processor` service is equivalent to a "yield" request indicating that the executing process is willing to yield to another process of equal or higher priority. If a user process is made ready by an interrupt handler during the null process's execution, then the yield request will cause a process switch to the new ready process with the null process returning to the ready queue. If no other user process is ready, then the null process will resume execution and continue in its loop.

In a preemptive system, the null process could be reduced to the following:

```
null_process( )
{
       while ( true )  {
       }
}      //end of null_process
```

The call to `release_processor( )` is not required since if an interrupt handler makes a user process of greater priority ready, then the interrupt handler will also perform a process switch to the higher priority process and place the null process back onto the ready queue. This will also improve the response time to switch to the user process that was made ready by the interrupt handler.

The null process could also perform some useful function, such as a diagnostic function, for example to compare checksums in ROM, etc. The first restriction for a null process is that it makes no API calls that could block it. In a non-preemptive system, a further restriction is that time duration between successive calls to `release_processor()` be short enough to allow an acceptable response to external interrupt generating events. This last additional restriction is not needed in a preemptive system.

6.4.3 The *release_processor* Primitive

The null process requirement led to the need for a `release_processor` primitive to be added to the RTX. A user process calls the `release_processor` primitive to voluntarily yield to another process (i.e. the currently executing process is requesting the kernel to switch to another ready process of equal or higher priority). If there is no such ready process, then the calling process will continue to execute. The basic pseudocode for `release_processor` can be given as:

```
K_release_processor:
       {
              set current_process state to ready
              rpq_enqueue(current_process) //put it on its ready queue
              process_switch()             //select next process and start it
       }
```

`K_release_processor` just puts the currently executing process back onto the ready queue, then picks the next ready process and performs a process switch to the selected process. If there are no other ready processes, then the caller gets picks again and continues to execute. The pseudocode could be optimized to reduce any latency.

6.5 System Initialization

At OS start up (e.g. power up reset), many initialization operations need to be performed such as hardware initialization, OS data structure construction, process creation, selecting the first process to run. In a dedicated embedded application, the OS must know which and how many processes to create. Our design will include an array of records (known as the Initialization Table (IT)). Each record entry contains all the information necessary to start its respective process. A record, for process_i, could have the following simple structure:

Record_i

process_id
process priority
required stack size
initial PC

During initialization, the IT would be processed and the following operations would be performed for each table record (process_i defined):

- allocate a new process control block (PCB) for the process.
- initialize the PCB with the process_id, priority and initial PC obtained from the IT record.
- allocate a block of memory from the stack space based on the requested stack size and place the value of the process's SP into the PCB.
- enter the process's initial PC into the PCB (this is the start address of the process code)
- since the process has never executed before, there is no previous context for this process. Create an initial context based on the SP and initial PC and save it into the PCB or to some other context save area for this process. Recall that a context includes the values of all CPU registers, status register (SR), etc. This initial context will be used with a 'RTE' (return from exception) instruction to start the process for the first time. Hence a previous context must be created for the RTE instruction.
- Set the state field of the PCB to “ready”, as appropriate.
- enqueue the PCB onto the ready queue in its proper position based on its priority.

As discussed in a later section, *iprocesses* will be handled with a slight difference in regard to the process status field during initialization. Once the entire IT has been processed, all internal kernel data structures created and all hardware initialized then the scheduler is invoked to pick the highest priority ready process and this process is started into execution.

7 IPC Primitives

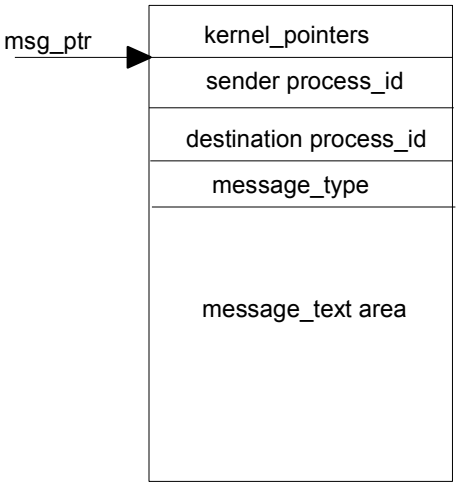
The design of several basic primitives is given. These primitives are only a sample of the required primitives and give an example of the design constraints and possible solutions. They are based on a basic non-preemptive system with some insight on the conversion to a preemptive system.

7.1 Inter-Process Communication (IPC)

In general, processes do not exist in isolation, they need to communicate with other processes and to synchronize their actions. The requirements call for a message-based IPC scheme. Messages are carried in shared memory blocks. A process writes a message into a shared memory block, sends a pointer to the memory block to another process and the receiving process reads the message from the memory block. The shared memory blocks are

called “message envelopes”. The sender of a message would invoke the RTX send message primitive supplying a pointer to the message envelope and the process id of the recipient. The immediate issues are: what is the format of the message envelope and where do message envelopes come from.

Message envelopes are managed by the kernel. To avoid run-time overheads, some fixed number of them are created during system initialization. Basically, this is a very simple form of memory management of fixed length memory blocks. A process would allocate a message envelope to send a message and deallocate an envelope when no longer required. The assumption is that a process owns a message envelope that it allocates or receives until it sends or deallocates the message envelope.

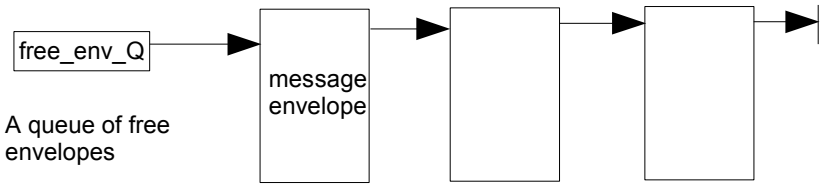


First consider the design of a message envelope and its internal structure. The adjacent figure illustrates a message envelope as a fixed length memory block with a default number of fields. The `kernel_pointers` field contains one or more pointers used by the kernel to place the message envelope on various linked lists. Ideally, this field should not be accessible to a user process. The next two fields hold the process_id of the sending process and the process_id of the destination (receiving) process. Typically, the kernel would update these fields whenever a message envelope is sent to another process. The `message_type` is an optional field which can indicate the type or class of the message to the receiving process. The actual text of the message is written into the message_text area by the sender. A receiving process can examine the sender process_id field to determine the sender of the message. Additional fields can be added as required.

7.2 Message Envelope Management

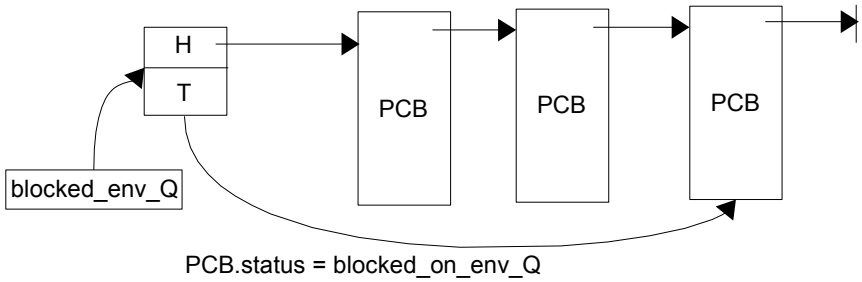
The question is where do message envelopes come from. We assume that a fixed number of message envelopes are created at initialization. This prevents any run-time overhead of allocating blocks of memory, to be used as message envelopes, on demand.

The fixed number of message envelopes are initially placed on a free envelope queue as illustrated. The kernel_pointers field is used to form a linked list of unallocated message envelopes.



The kernel uses the variable `free_env_Q` to hold a pointer to the first free envelope. If a process does not have an envelope then it allocates an envelope first. If there are no free envelopes available, then the process blocks and must wait until an envelope becomes available (i.e. deallocated by some other process). The blocked process can be placed on a 'blocked for envelope queue' joining any other processes waiting for a free envelope as shown in the following figure. The PCB of the blocked process is enqueued onto the blocked queue either in FIFO order or the queue can be sorted according to process priority (as appropriate). When a message envelope becomes available, then a PCB can be dequeued from the blocked queue and be given the available message.

7.3 Message Envelope Allocate



To obtain a message envelope, a process calls the `allocate_envelope` API primitive. If there is a free message envelope, then the process is returned a pointer to the free envelope. If there is no free envelope available, the process blocks and is placed on the blocked envelope queue and a process switch occurs. As an example, consider the following pseudocode for allocate_envelope (where MSG is a type specifier for the message envelope):

```
MSG * K_allocate_envelope() {

       while (free_env_Q is empty) {
              put PCB on blocked_env_Q
              set process state to blocked_on_env_allocate
              process_switch();  //the caller blocks
                     // restart here when blocked process executes eventually
          }
       env ← reference to de-queued envelope
       return env;
}
```

The “while” loop is used to allow the kernel to retry its check on the availability of a free envelope even if the process becomes ready. For example consider the following scenario. A process blocks on its attempt to obtain an envelope. Sometime later a different process deallocates an envelope, and the blocked process is made ready and placed on the ready process queue before it can try again (due to the process switch). While it is on the ready queue, another executing process might possibly allocate the currently free envelope since it was not directly given to the blocked process. Then when the process attempts to check the free envelope queue, it might find it empty causing it to become blocked once again, and so on. As will be seen later, this sequence can be optimized once the deallocate envelope primitive is discussed. In this example, the allocate_envelope primitive only makes a blocked process ready, put on the ready queue and is given the opportunity to try again to obtain an envelope. There is no guarantee that a free envelope is still available once the blocked process restarts after the call to process switch. So, the primitive works, but could be made better.

7.4 Message Envelope Deallocate

A process calls `deallocate_envelope` to return an envelope it no longer requires. The freed envelope is placed on the free_envelope_q. The blocked queue of processes waiting for envelopes is checked. If the queue is empty, then the primitive finishes. Otherwise, a blocked process is removed form the blocked queue, made ready, placed on the ready process queue allowing the process to retry its allocation again. The following pseudocode, for a non-preemptive system, illustrates a possible implementation:


```
int K_deallocate_envelope( MSG * env ) : {

       put env onto free_env_Q
       if ( blocked_env_Q not empty)
         { dequeue one of the blocked processes
            set its state to ready and enqueue it on ready process queue
          }

}
```

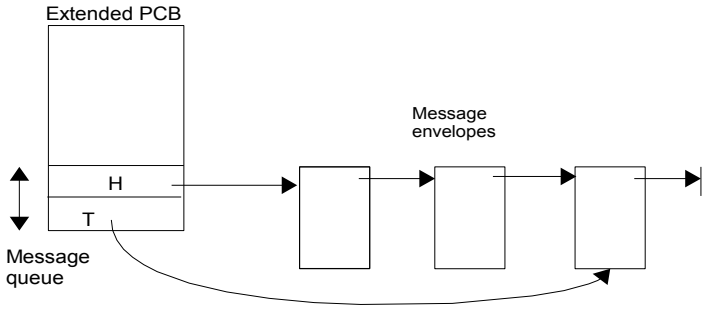
The primitive does not block the calling process. The freed envelope is not directly given to a waiting process, but rather the process is made ready and allowed to retry its allocation attempt. This can be improved by providing a mechanism to give the freed envelope directly to the process made ready and then modifying the allocate primitive accordingly. This is left as an exercise.

Although the IPC primitives could be improved to immediately deliver the freed message envelope to a waiting process, consider the following scenarios where we assume a preemptive system and the blocked on envelope queue is sorted by process priority:

1. If deallocate envelope makes a waiting process ready and the ready'd process is of higher priority than the process performing the deallocate, then an immediate process switch will occur and by default the newly ready'd process will obtain the envelope.
2. If deallocate envelope makes a waiting process ready and the ready'd process is of lower or equal priority than the process performing the deallocate, the ready'd process will wait on the ready queue until it is scheduled. No process of lower priority can allocate the newly freed envelope. But since the ready queue is FIFO ordered when processes are of equal priority, another process (ahead of it in the in the queue) of equal priority could allocate the newly freed envelope. This makes the primitive implementation seem less fair since the original process might possibly block a second time when it is scheduled.
3. If deallocate envelope makes a waiting process ready and the ready'd process is of lower or equal priority than the process performing the deallocate, the ready'd process will wait on the ready queue until it is scheduled (a further continuation of scenario 2). Next, a higher priority process is either scheduled to run or made runnable by an i_process. The higher priority process can now allocate the newly freed envelope before the original process. This may seen unfair to the original process but it does follow the process priority objectives and may even reduce effects of priority inversion.

7.5 Message Receive

The requirements call for a synchronous message receive. An initial design issue is how does the kernel buffer messages that are sent to a process but the process does not do a receive for some time. A simple solution is to have the kernel maintain a message queue for each process and place the queue head/tail structure in the process's PCB as an extra field. The following figure illustrates how the PCB is extended to perform this additional functionality. Each PCB now has a new message queue field to hold messages sent to a process but



not yet received.

The queue could be implemented as either FCFS or be sorted based on the priority of the sending processes.

If a process does a receive and there are no messages waiting for it, then the process blocks with its status set to blocked_on_receive.

The following pseudo code illustrates a possible implementation of the receive message primitive:

```
MSG * K_receive() {

   if ( current_process's msg_queue is empty) {
         set state of current_process to blocked_on_receive
         process_switch( );
       *** return here when this process executes again
   }
   env ← dequeued envelope from the process' message queue
   return env

}
```

If the process's message queue is empty, then the process blocks. Eventually the process will restart once a message is sent to it. At that time, the process is returned a pointer to a message envelope.

7.6 Message Send

Since the message send is asynchronous, the sending process does not block on a message send. The message is delivered to the message queue of the destination process. If the destination process is currently blocked waiting for a message, then the process is made ready and placed on the ready process queue. The following pseudo code is an example of a possible non-preemptive implementation:

```
int K_send( int target_pid, MSG *env) : {

   set sender_procid, destination_procid fields in env
   target_proc ← convert target_pid to process PCB reference
   enqueue env onto the msg_queue of target_proc
   if ( target_proc.state is blocked_on_receive)
     { set target_proc state to ready
       rpq_enqueue( target_proc );
     }
```



```
}
```

7.7 Possible Extensions

Although the original requirements did not specify other forms of send/receive, it is relatively easy to add additional forms. For example an asynchronous receive can be added to return, say a null pointer, if there is no message waiting for a process (this is done for the special case of an i_process calling receive). Also, a synchronous send can be added to block a sender of a message if the destination process is not already blocked waiting for a message and then make the sender ready when the message is finally given to the destination process when it invokes a receive.

In addition, we could add a receive which specifies the expected sender of the message. Here, several semantics are possible. First, a null message pointer could be returned if a message from the specified sender is not already waiting. An alternative could be to give 'priority' to messages of the specified sending process if there are multiple waiting messages.

7.7.1 Prioritized Message Queues

Messages waiting on a process's message queue could be enqueued either in simple FIFO order or be sorted according to priority. If it is chosen to sort the messages according to priority, then the next design problem would be the definition of 'priority' from the perspective of the message envelope. Does a message inherit the priority of the sending process or is the priority encoded into the message itself regardless of the priority of the sender. Either definition (or both) could be used depending on the application.

7.7.1.1 Inherited Message Priority

Consider the case of a message inheriting its priority from the sending process. This could be valid when a high priority process sends a message to a lower priority process. Thus when the lower priority is eventually scheduled, it is assured that it will receive messages from any higher priority processes before any other messages are received. This is also valid when high priority processes send messages to each other. One obvious consequence is a possibility of starvation of the lower priority messages (from low priority processes) if the higher priority processes are very active in sending messages to each other. If the message traffic is mostly between processes of 'equal' priority, then the inheritance based priority sorted queue is of little advantage.

Another design issue is how its priority is maintained by the message while enqueued on the message queue. Its priority will need to be examined each time a new message is enqueued. Currently, the message envelope has no capability to maintain priority information. This can easily be fixed by adding another field to the message envelope called, say '`msg_priority`' . The modified send primitive would copy the sender's priority (from its PCB) into the message envelope's *msg_priority* field prior to placing the message envelope into the priority sorted message queue of the destination process. This would be a permanent addition to the fields of a message envelope.

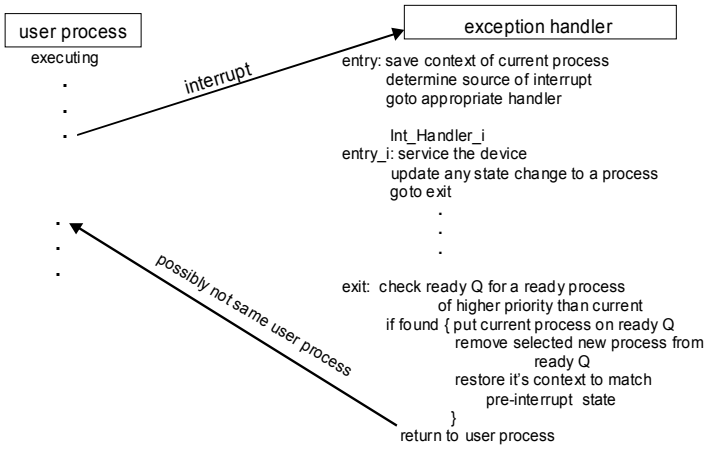
7.7.1.2 Priority based on *message type*

It might be possible to imply the priority of a message based on its message type. The problem is that in many situations, the message type has only meaning (or significance) to the processes exchanging messages of a particular set of message types. Thus, the kernel (or send primitive) would need to be aware of the sets of communicating processes and understand the meaning of the message types in order to infer priority. This could be done by defining a fixed set of global message types in some priority order. But this does not seem to be very

attractive or desirable. If done, then sorting of the messages can easily be performed by examining the existing message type field of the message envelopes.

8 Interrupt Handling

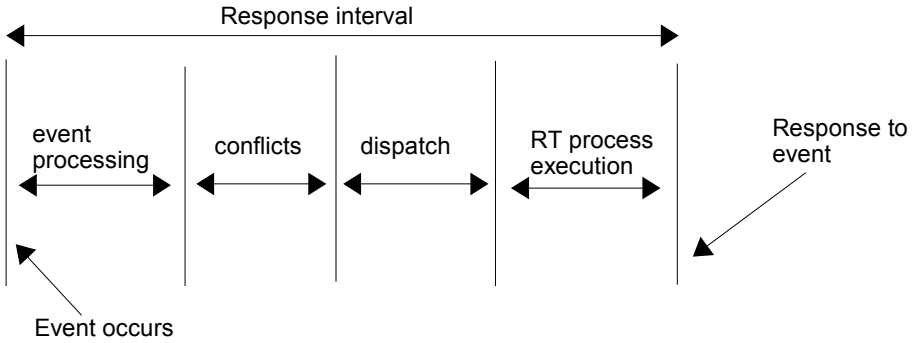
In a real-time system, interrupt handling is one of the more important design issues since the latency in response to an interrupt must be at a minimal and requires efficient and speedy interrupt handlers. Interrupt processing may cause a blocked process to be made ready and placed on the ready process queue. For example, a process blocked for an external event to occur will have its state changed to ready and be placed on the ready process queue when the event occurs. In a preemptive system, the user process executing at the time of the interrupt could be preempted by a higher priority process made ready by the interrupt handling. A possible interpretation of an interrupt is a hardware message requiring a short response latency and short service time.



Design issues include: does the interrupt handling code run as part of the kernel, within a process (if yes, which?); are interrupt handlers themselves interruptible; if the OS is preemptive, need to deal with the possibility that an interrupt results in a higher priority process becoming ready and performing an unscheduled process switch.

The adjacent figure illustrates a generic interrupt or exception handler operation for a preemptive system. After the servicing the interrupt, the exception handler checks to see if a higher priority process was made ready by the service. If yes, then a process switch is made to the newly ready'd process and the original process running at the time of the interrupt is placed on the ready process queue. The figure depicts the situation where all interrupt sources (devices) are 'OR-tied' to a single CPU interrupt line. Thus, the initial interrupt handler determines the device that caused the interrupt and then transfers control to the specific handler for the device.

One of the more important design issues involves the latency (or the elapsed time) between the occurrence of the interrupt and when the real-time process that is waiting for the interrupt is actually run and has performed the necessary actions related with the interrupt. The following figure illustrates the situation.



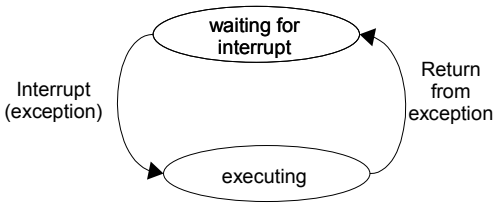
When the event occurs, it requires some initial event processing to make the RT process ready and place it on the ready process queue for scheduling. There may be conflicts since it may have to wait its turn before being chosen for execution. There is some dispatch latency to select the process, restore its context and start its

execution. Finally, the RT process needs execution time to complete the event processing. One goal is to make the response interval as short as possible by reducing the event processing, conflicts and dispatching overheads.

8.1 Concept of an *i_process*

In general, the interrupt handler must interact with the real-time process that will be performing the event handling as discussed in the last section. There are many ad-hoc mechanisms that could be employed to reduce the response interval. Let us consider a structured approach using the concept of an *i_process*.

The i_process gets the CPU through an interrupt handling sequence, not through the dispatcher. It never blocks if it invokes a kernel primitive (the primitives must be aware that an i_process invoked the primitive). The interrupt (exception) handling routine starts the appropriate i-process. Conceptually, the i_process has the highest priority and is scheduled (started) by the interrupt handler. The state diagram for an i_process is shown. The i_process can execute in kernel mode which may give it special privileges not available to the basic user level process.



i_process state diagram

When idle, the i_process is always ready to run but not placed on any ready queue. The i_process has an associated PCB with the state field permanently set to "`i_process`". An i_process can invoke any kernel primitive, however, an i_process is not allowed to block! Primitives which can block a process must be modified to ensure that an i_process does not block! For example, the synchronous receive message primitive should be modified to return *null* if the invoking process is an i_process and there is no message waiting. Similarly for other primitives which can potentially block a process or perform a process switch. If the design of an i_process requires that it send messages during its execution then it must be ensured that the i_process has a sufficient supply of envelopes or can get them without invoking the allocate_envelope primitive. In general, an i_process should not attempt to allocate resources during execution since there is a real probability that there will not be any resources available. Also, any kernel primitive which can potentially perform a process switch should check to ensure that it does not attempt to perform a process switch on an i_process.

The i_process could be considered as a high level "intelligent" handler of an interrupt. Thus, it might be desirable for the i_process to maintain state information between interrupts (executions). The design of each i_process should take this into account. Some issues include: is it required to perform a complete context switch to run an i_process or can it be invoked as a simple C-function (in either case a PCB would be required); how would state information (memory) be maintained across invocations; what additional private data structures would be needed by each i_process, etc..

An i_process has some privileges that a user process does not. For example, it can call directly the kernel version of the user visible API without calling the user visible API functions and entering the kernel by a SWI. It is already in the kernel due to an interrupt handling sequence.

8.1.1 Example

As an example, consider the case where a user process sends a message to a UART-i_process. The message text contains a string of characters to be send to the UART. The i_process delivers each character to the UART one-at-a-time on each TX interrupt from the UART. On each TX interrupt from the UART, the interrupt handler starts the respective i_process. The i_process does a receive to see if there are any new messages sent to it, and if yes, places them on some private service queue. It then checks its previous state to see which character (if any) it must send to the UART from the current message envelope on its service queue. If the last character has been sent from the current message envelope, the message envelope is sent back to the requesting process as an

acknowledgement. Next, the service queue would be checked to see if more messages are waiting for service. When the last character of the last message is sent, then a state variable is set to indicate that the i_process is idle and that the UART is in a ready state since no further TX interrupts would be caused by the UART until it is loaded with another character.

8.1.2 Possible Complications

The previous section described an i_process sending characters to a UART in response to TX interrupts and an overview of its functionality. Generally, a UART will only give one interrupt when its transmit buffer is empty. Usually, this is sufficient for the i_process to load the next character to the UART. What happens when the i_process has sent the last character in the last message on its service queue and it has returned the message envelope. Potentially, the i_process will never run again since the TX interrupt handler will not run because there will be not any more TX interrupts until the transmit buffer is loaded again. So, if messages are sent to the i_process, the i_process will never receive them and the characters will not be sent to the UART since the i_process must be started by the interrupt handler which will not run. So, we have a problem. The same situation occurs after initialization and the first message is sent to the i_process. A similar scenario may be possible for other i_processes defined in the system. This type of problem does not occur for periodically occurring interrupts and their associated i_processes. What we need is to either avoid this situation (?) or perhaps have a way to "jump-start" an i_process for the first time (or periodically). The possible solutions are left as an exercise.

8.2 A generic interrupt handler

We can now update the previous generic exception handling sequence to, assuming non preemptive for now, with the following pseudocode:

exception_handler:

```
begin
       // code here to save current_process context into its
       //context save area as defined by its PCB
       save_PCB = current_process
       select interrupt source
              A:     current_process ← i_proc_A_pcb
                     //restore i_proc_A context
                     //invoke i_proc_A handler
                     break
                     ........
              Z:     current_process ← i_proc_Z_pcb
                            //restore i_proc_Z context
                            //invoke i_proc_Z handler
                     break
                     end select

              //code to save context of interrupt handler (i_process)
                     current_process = save_PCB;

              //code to restore current_process context
              //
              //perform a return from exception sequence
              //this restarts the original process before i_handler
              end;
```

For a preemptive system, a check would be made on the return from the i_process to determine if the ready queue contains a ready process of higher priority than the process that was executing at the interrupt event. If yes, then the original process would be preempted by the process made ready by the i_process causing an unscheduled process switch.

There is another design issue to consider. During its execution, an i_process may need to invoke a kernel primitive. Does the i_process use the same user API interface that a user process uses, or does it directly call the kernel primitive? Recall that the i_process executes during interrupt handling within kernel mode. Thus it could bypass the user visible API interface and directly call the kernel version of the primitive avoiding any overhead and complexity of re-entering the kernel with a TRAP instruction. Since kernel primitives will not block an i_process, there is no danger of the i_process becoming blocked within the kernel and requiring a process/context switch to be performed.

8.3 Timing Services

This is one of the more fundamental required services. Real-time processes interact with their environment and need services such as:

- delay : delay my execution for *n* seconds. Voluntarily give up execution until the specified time expires, then put back on ready queue (often referred to as sleep).
- timeout : requests kernel to inform process when a specified time period has expired, the process continues execution.
- repetitive timeout : request to be informed every *n* microseconds until cancelled

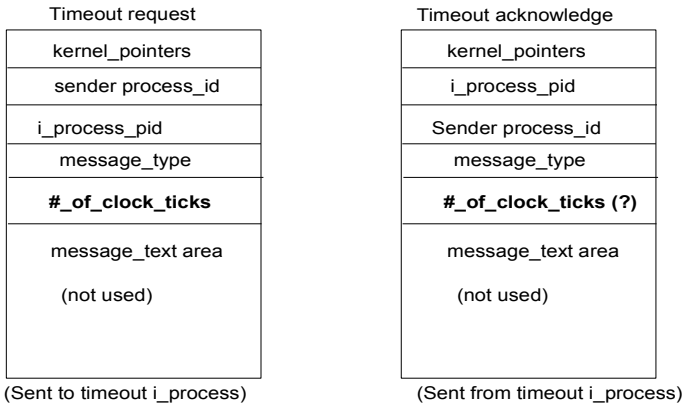
These requests could be in:

- relative time : *n* number of clock ticks from now
- absolute time : for example, February 24, 10:34:22 AM, 2008

as appropriate. "*clock ticks*" are defined as the number of interrupts from a hardware device which periodically sends an interrupt to the processor. For example, the processor could receive an interrupt every 10 ms. In that case a "clock tick" would be 10 ms in duration.

The basic design issues include: what is the protocol to access the timing services; "who" performs the services; and how is the invoking process informed of the timing service's results? Since we assumed a message passing RTX, we can construct the timing services within that framework. We already have the concept of an i_process. Let a user process send a message to a special i_process which performs the service. The i_process then sends a message back to the requesting user process as an acknowledgement.

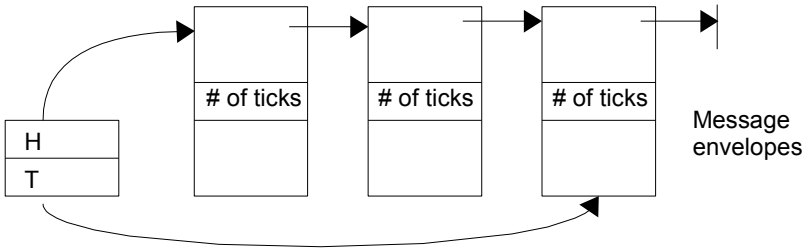
Since message passing is the basic form of IPC, we can add an additional field to the message envelope structure when it is used to communicate with a timing service i_process. Consider a simple timeout service accessed by the standard `send(target_pid, message_ptr)` primitive. The adjacent figure depicts the resulting message envelope with a new field called "# of clock ticks".



All user processes 'know' the pid of the timeout i_process. After the expiration of the time, the timeout i_process sends the original message envelope back to requester. The timeout service maintains requests in a private ordered queue. There are many possibilities in the type of queue and the insertion of timing requests into the queue. A design goal would be to i) minimize the overheads involved for the insertion of messages onto the queue and ii) to determine that the timeout interval has expired for any individual request.

At each clock tick (hardware timer interrupt), the timeout i-process executes. The i-process invokes the receive message primitive to check for any new requests since the last clock tick interrupt. Since it is an i_process, it cannot block if there are no messages waiting for it. If there are any new requests, then they are added to the sorted list. Then the i-process checks if any timing requests have expired. If yes, then remove the message(s) from the sorted timeout list and then send the message envelopes back to requesters as an acknowledgement of the timeout expiration.

There are many possibilities of how to maintain a sorted timeout list. As an example, consider the figure below which depicts one such sorted list. The # of ticks field can be modified by timeout to indicate the # of ticks after



Timeout queue

its predecessor in the list. New timeout requests would be inserted in sorted relative timeout order. As an example consider a timeout queue with the following # of ticks values:

- queue contents of timeout list = {25, 30, 0, 10} (the # of ticks field shown)
  - one timeout for 25 clock ticks
  - two timeouts for 55 clock ticks
  - one timeout for 65 clock ticks

Each request must be inserted into the appropriate location of the queue and all requests which follow it in the queue might also have their clock tick adjusted. After insertion, checking for a timeout expiry requires only to decrement the request at the head of the queue (see the following pseudocode). This method requires some overhead to install the timeout request onto the timeout queue (done once per request) but can very efficiently determine when a timeout expires.

In addition, the timeout i_process can also maintain the internal kernel clock. This clock is based on clock ticks and is an integer (i.e. 32-bit) counter which is simply incremented at each clock tick interrupt. This clock can also be used to perform the timeout operation. For example, for each timeout request received, add the specified number of clock ticks to the current kernel clock and store this value as the timeout expiry time in the received message envelope. The timeout queue would be sorted using this field. Then, on each interrupt simply compare the expiry time of the head of the timeout queue with the incremented kernel clock. A timeout has expired when the timeout request is less than or equal to the kernel clock.

The following is a basic outline of service execution for an interrupt based on maintaining timing requests in a relative sorted order as discussed previously:

```
timeout_i_process:
{
        env ← K_receive(); //invoke primitive to get pending msgs
        while (env is not null)
        {
                //code to sort the envelope into the timeout_list
                env ←K_receive(); //see if any more msgs left
```



```
        }
        if (timeout_list is not empty)
        {
                //code to decrement the tick count of 1st in the timeout_list
                while (head_of_timeout_list[tick_field] is zero)
                {
                        env ← dequeue( timeout_list );
                        temp ← env[source_pid];
                        env[source_pid] ← timeout_i_process_pid;
                        K_send( env, temp );    //return envelope
                }
         }
}
```

9 Process Preemption

For the system described, preemption is simple to add. For example each time a process is placed on a ready queue by some primitive (e.g. possibly by a send, deallocate memory or message envelope, or by the operation of an i_process) perform the following:

- check if the ready queue contains a ready process whose priority is higher than the currently executing process,
- if yes, then perform a process switch (at appropriate time) to the higher priority process.

A check would be made by the interrupt handler just before control is returned to the process executing at the occurrence of the interrupt. Thus a process switch could be made just before the RTE of the interrupt handler.

Key design issues are “at what point is the process context saved?” and “at what point is a process switch made?”. For example, when a user process invokes a kernel primitive, is the actual process context saved by the TRAP handler before the kernel primitive is called by the trap handler? Or is the complete process context saved by the context switch function if there is a process switch. Which stack pointer is used after entry into the kernel by a SWI instruction: the user's stack pointer or do we switch to a special kernel stack pointer?

10 Process Switching – Context Switching

Performing a process switch requires a more detailed examination. A process switch can be performed at the SWI handler boundary (after returning from a kernel primitive) or within a kernel primitive itself. In this example, we allow a process switch to occur within a kernel primitive as needed. Recall that a process switch will be requested either by a kernel primitive while processing a user request or by the interrupt handler when a process preemption is required due to the action of an i_process. Lets examine the case of a process switch requested by a kernel primitive and as an example consider the receive message primitive which is repeated below:

```
K_receive() { //need a check for i_processes

   if ( current_process's msg_queue is empty) {
         set state of current_process to blocked_on_receive
         process_switch( );
       //*** return here when this process executes again
   }
  env ← dequeued envelope from the process' message queue
  return_val env
```



```
}
```

In this example, the current process will block inside the kernel and will eventually restart execution within the kernel as well. After the process restarts, it will return to the next line of code following the call to process_switch(). The net effect is that of a simple call to some function called process_switch() which eventually returned and the kernel primitive continues execution without any knowledge that the process was blocked for some time interval. Hence, all the "complexity" is within the function process_switch().

Process_switch is not informed of the process requesting a process switch since the kernel variable `current_process` is pointing to the PCB of the process that invoked receive in the first place. What functionality does process switch have? At a minimum, in this example, it will call `rpq_dequeue()` to obtain a pointer to the PCB of the highest priority ready process (the process which will execute next). Thus the simplest basic skeleton of process switch could be:

```
process_switch( )
{ PCB_ptr *next_pcb;          //local variable to hold ptr to next pcb
  next_pcb = rpq_dequeue( );  //get ptr to highest priority ready process
 context_switch( current_process, next_pcb );
}
```

Process_switch does not enqueue the `current_process` 's PCB on any queue nor modify the status field – this must be done by the caller of process_switch. Process_switch could change the status field of next's PCB to executing. Thus, the current process will be "stopped" inside of "context_switch()". The next process will restart inside of context_switch(), then return to process_switch and then return to whoever called process_switch (assuming that the next process was also stopped in this manner). Fundamentally, when a process restarts inside of context_switch, it will return to what ever function called context_switch. Care should be done in context_switch's design to made it "bullet-proof" and allow it to be called by any function (not just process_switch).

The next issue is: what is the design of "context_switch"? At this point, context_switch has some fundamental requirements:

1. Save the context of the current process into its PCB save area allowing it to restart such that a return will be made to the caller of context_switch. This includes, the stack pointer, PC and SR.
2. Restore the context of the next process (from its PCB save area including the SP, PC and SR)).
3. Possibly set the status field of the next's PCB to executing.
4. Possibly set the kernel variable `current_process` to point to next's PCB.
5. Restart the next process such that it "returns" to who ever called context_switch.

The structure of context_switch is:

```
context_switch( pcb_ptr *current, pcb_ptr  *next) {
    char * temp1;  //some temporary variables
    char * temp2;
    // code to save context of 'current' process
    // possibly set next->status = executing
    // possibly set "current_process" to point to next's PCB.
    // code to restore context of 'next' process
    ASM("RTE");              // return from exception
```



```
Comehere: ASM("nop");
          ASM("nop");
          return;    //normal C-language return statement
}
```

What is the value of "PC" that will be saved in the PCB of "current"? When 'current' restarts, we want it to return to the caller of context_switch. This is easy, we actually want it to restart at the label "`Comehere:`". Therefore, when the current process eventually restarts, it will execute the "`nop`" instructions and then hit the return of `context_switch` and thus returning to `process_switch` (or whoever called it) using the normal C-language conventions. So we need to add code to get this value (the address of `Comehere:` label) and save it into current process's PCB field for PC. The RTE instruction requires a specific sequence of data on the stack (pointed to by register a7 -the stack pointer) prior to the execution of the RTE. Look in the MCF5307 programmer's manual for any details. Basically, the following sequence is performed:

$$2 + (SP) \rightarrow SR; \quad 4 + (SP) \rightarrow PC; \quad SP + 8 \rightarrow SP$$

The value of PC and SR are obtained from next's PCB save area and are pushed onto the stack last in the proper order. In addition, the Format/Vector fields must contain an appropriate bit sequence. You can push these onto the stack with the SR value. Unless you want to become a "guru" MCF5307 assembly language programmer, avoid pushing any 8 or 16 bit values onto the stack in your program. Always maintain a long-word aligned stack pointer.

With some thought, we might be able to use the basic ideas here to perform a process preemption handled by an interrupt handler after an i_process makes a high priority process ready; perhaps even using a 'bullet-proof' version of the context_switch just discussed.

11 A zero security minded RTX?

One of the assumptions for the kernel design was based on user processes that were friendly, cooperative and always did the 'right thing' and were non-malicious. This assumption 'might' be valid in the final system but may be optimistic during the initial stages during the testing and debug phases of user processes. In the current design, it is rather easy to crash the kernel due to a bug in a user process or by a malicious action on the part of a user process. Perhaps the first line of defence would be the hardware design of the embedded microprocessor system itself. The hardware should make use of processor generated output signals indicating user mode and supervisory mode execution to ensure that no kernel address space is accessed by a user process while the processor is in user mode execution. Such attempted access should generate an exception condition forcing the processor to perform exception handling with kernel supplied exception handlers. Also, if available, virtual memory can play a valuable role. The virtual to physical memory page table translations can ensure a separation of address spaces between all user processes and the kernel address space preventing any unwanted interaction. Further discussion on these points is beyond the intended scope of this document (discussed in lectures).

The current design can be upgraded to provide a minimal level of security during initial system debug (with minimal performance impact) and then be disabled (??) when the system proves to be stable. First, consider the primitives `send`, `receive` and `deallocate_envelope` . For `receive`, the user process assumes that the kernel has provided a valid pointer to a message envelope. For `send` and `deallocate_envelope`, the kernel trusts that the user process is providing a valid pointer to some kernel generated (during initialization) envelope (block of memory). The kernel does no check to verify the validity of the supplied envelope pointer. If this pointer is invalid, then a crash of the system may be inevitable. Also, the kernel does not track ownership of message envelopes after they are allocated. There is nothing to prevent a user process from deallocating an envelope after it has sent the envelope to another process which could result in a system instability. A process could write into the envelope a message text which overflows the length of the envelope and possibly overwrite another

envelope. Ideally, these problems 'should' not arise in the final production system, but could occur during the debug phase.

The kernel can easily place all envelopes (created during initialization) on a linked list using the kernel pointers field within each message envelope and then scan the list to verify a valid pointer is passed to the kernel. But a user process most likely also has access to these fields and could easily change them causing instability within the kernel since theses fields are also used to place free envelopes on the envelope free list. This simple approach adds some level of security but can be defeated by an errant process.

Assume there is at least some separation of kernel and user address spaces provided by the hardware design of the controller. This would prevent wholesale invasion of a user process into the kernel's private data structures. There are many possible approaches to verify that a valid envelope pointer (which a process currently owns) is provided by a user process during some primitive invocation. The following describes a simple approach that can provide a level of security. This scheme can easily be optimized and left as an exercise.

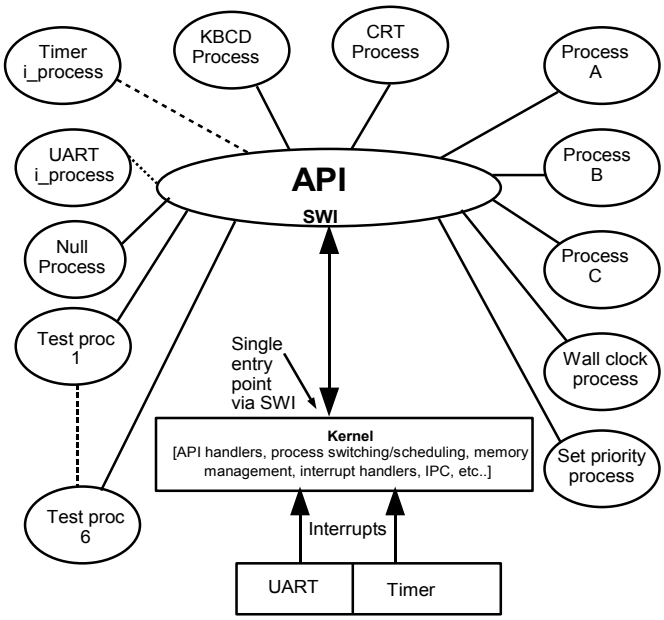
We can define an envelope control block (ECB) similar in structure but simpler than a PCB. During system initialization, one ECB is created for each envelope memory block. It is the ECB, not the message envelope, that is placed on the envelope free queue. The message envelope's structure is still the same with all fields except it is missing the kernel pointers field since it is not needed. The kernel also maintains a linked list of all ECBs in the system. The ECB contains (at a minimum) a kernel pointers field (used to place it on various linked lists), a pointer to the actual memory block of the envelope the ECB represents and a field containing the PID of the current owner of the envelope.

On a message send, the send primitive would copy the destination process's PID into the current owner field of the ECB. The allocate envelope primitive would copy the requesting process's PID into the current owner field of the ECB when an envelope is allocated; and obtain the pointer to the envelope from the appropriate field within the ECB. The deallocate envelope primitive would reset the current owner field of the ECB indicating that the kernel is the owner. Thus any attempt to send or deallocate a message envelope not owned by the requesting process can be easily detected by the kernel's primitives. Also, any attempt to deallocate/send an envelope by supplying an illegal envelope pointer can be easily detected. Depending on the implementation, detecting an illegal envelope pointer is perhaps the largest overhead without a consistent execution time. The overheads of maintaining the current owner field in the ECB is negligible.

Other simple extensions (causing minor revisions to existing primitives) can include the verification that the destination process PID in a message send is valid, etc.

12 Appendix A: User API

This section discusses the user API from a generic point of view with an example using the MCF5307 processor. The following figure depicts a simplified view of the overall process-kernel structure of a typical application. You might have additional processes in your design. Processes access RTX services via a single SWI (Software



Interrupt) from the specified user visible API. The i_processes are shown with dotted lines since they run in kernel mode so they might access the kernel version of the API functions directly rather than following the user process protocol. In a typical system, user processes and user visible API functions can not access the protected address space of the kernel. None of the kernel functions or data structures are visible or accessible while executing in user address space. These are only available in the kernel's address space while the processor is executing in "supervisor mode". For example, the user visible/called API "receive_message" can not itself perform the requested service. Basically, it passes the request to the kernel by executing a software interrupt instruction (a SWI or TRAP instruction on most processors). This causes the processor to begin an exception handling sequence which starts the kernel's trap handler (in the kernel address space and in the processor's supervisor mode of execution).

12.1 Example of a single kernel entry point design.

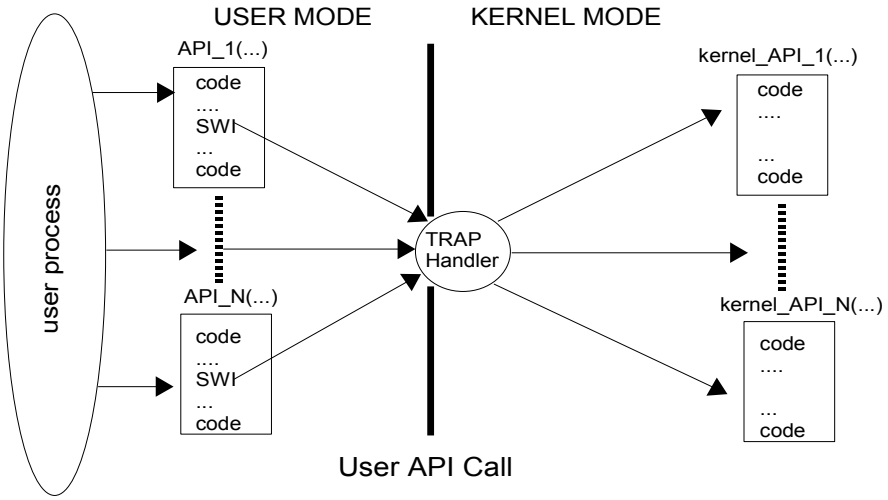
All entries to kernel are by the same SWI (a TRAP in the MCF5307) instruction in the user process visible API_i(..) function. All API_i(..) functions invoke the same SWI instruction after presetting various CPU registers to indicate to the Kernel_API_i(...) of the functionality required plus any additional parameters obtained from the stack frame of the invoking process.
The SWI trap handler (within the kernel) would examine the CPU registers and determine which Kernel_API_i(..) function to invoke. The trap handler would save some/all of the process context (depending on the design) and ensure that interrupts are disabled during the kernel's execution. The trap handler would also

disable some (or all) interrupts ensuring atomicity while in the kernel.

Comments:
In general, the API functions (as described in the project document) should **not** access the private variables, data structures, queues, etc. of the kernel. Only the respective Kernel_API_i(..) functions have the privilege of doing so. The specified behaviour is performed by the Kernel_API_i(..) function not by the published API, since these only serve to provide a controlled entry point into the kernel.

The user visible API functions would perform (at least) the following:

1. for an API function call with *i* formal parameters, move registers d1 .. d*i* onto the stack (save them).
2. Copy the values of the actual parameters (from current stack frame) into registers d1 .. d*i*.
3. Load register d0 with an appropriate Kernel_API_*k* integer code (used to select which function to invoke).
4. Perform a software interrupt (SWI) using an appropriate TRAP #*n* instruction where *n* is the trap number.
5. The kernel's trap handler would determine which kernel_API function to invoke and pass to it the required parameters obtained from d1 .. d*i* registers; and then finally return to the user API by a return-from-interrupt instruction.
6. On return from the TRAP, register d0 should already be pre-set with the required return value to be returned to the user process.
7. Now, restore registers d1 .. d*i* previously saved on the current stack frame.
8. Use the gcc supplied return mechanism to return control back to the original caller.



12.2 An example for the generic RTX

As an example of where this information can be used, consider the project's user process visible send message API function:

`int send_message (int process_ID, char * MessageEnvelope)`

In our implementation, `send_message` is just a skeleton API function (a stub) with no real functionality except to pass the request to the kernel, which will then perform the send message functionality, and then return the results to the process which invoked `send_message`. The user visible `send_message` does not have the right (or ability) to disable interrupts nor access internal kernel data structures. The request is transferred to the kernel via a trap instruction. The kernel's trap handler examines the request code placed in register `d0` (after disabling interrupts and ....) and calls the respective kernel function supplying the parameters available in the `d1` and `d2` registers. In this example, the actual functionality of `send_message` would be in a C-function called:

`int kernel_send_message (int process_ID, char * MessageEnvelope)`

The trap handler would call `kernel_send_message`. Now, `kernel_send_message` runs in kernel mode and can access all kernel variables, queues, etc without any interference from interrupts and user processes.

Another item for consideration is: would an i_process call `send_message` or `kernel_send_message`? Recall that i_processes only run during interrupt handling, so.....

Below is an example (of many possible) implementation of the `send_message` API stub.

```
int send_message( int process_ID, char * MessageEnvelope)
{
   int retCode = 0;
   asm("move.l %d1, -(%a7)");    //save the d1 register on stack
   asm("move.l %d2, -(%a7)");    //save the d2 register on stack
   asm("move.l 8(%a6),%d1");     //copy 'process_ID' into d1 register
   asm("move.l 12(%a6), %d2");   // copy 'MessageEnvelope' into d2 register
   asm("move.l #7, %d0");        //assume that send_message has kernel_code of 7
   asm("TRAP #0");               //assume trap#0 is the kernel swi trap handler
                                 //we assume that the return value will be in reg. d0
   asm("move.l (%a7)+, %d2");   // restore reg. d2
   asm("move.l (%a7)+, %d1");    // restore reg. d1
   asm("move.l %d0, -4(%a6)");   //copy returned value to local variable 'retCode'
   return retCode;
}
```

Recall the MCF5307 trap instruction sequence:

```
1 → S-Bit of SR
SP – 4 → SP; nextPC → (SP); SP – 2 → SP;
SR → (SP); SP – 2 → SP; Format/Offset → (SP);
(VBR + 0x80 + 4*n) → PC
where n is the TRAP vector number
```

The trap handler would do an initial save of registers (some or all, depending on design) and then examine the d0 register to determine which kernel primitive is to be invoked. This could be done by a C-language switch statement after saving d0 into some local variable. Before calling the appropriate C-language kernel primitive, push the required arguments onto the stack (using assembly language statements) – these were in the d1, d2, d3 registers when the TRAP was executed and then call the kernel primitive using a "bsr" assembly language call (or place the contents of d1, d2, d3 into local variables and then using C-language call the kernel primitive with appropriate actual parameters). On return, the kernel primitives' return value will be in the d0 register (gcc default). Next re-align the stack by adding an offset to account for the passed parameters (still on the stack) – if you called the kernel primitive with a "bsr". Restore any previously saved registers (be sure to leave the new

value of d0 in d0) and execute a RTE instruction which should return to the origin user level API function that executed the TRAP instruction. See the section on Process/Context switching for details of a process switch occurring when a kernel primitive is called.

12.3 MTE 241 API Constraints

Since the MTE 241 project is implemented by designing an RTX on top of UNIX, the above example should be used to only appreciate the basic constraints and possible solutions for the general case of an embedded design using a dedicated processor. Within the project constraints (due to UNIX environment), the user processes, the kernel and its primitives and all interrupt (signal) handlers execute in the user mode of the processor. The special supervisor (or kernel) execution mode is not available. Thus, some of the basic RTX design must be 'moulded' to fit the environment.

The project description gives an example for a RTX class definition as:

```
class MsgEnv;

class RTX {

  private:
      // any required private methods and attributes not user visible
  public:    // only a subset of the public methods
   int  send_message( int dest_process_id, MsgEnv * msg_envelope );
   MsgEnv * receive_message( );
   MsgEnv * request_msg_env( );
   int  release_ msg_env ( MsgEnv * memory_block );
   int  release_processor(  );
   int  request_process_status( MsgEnv * memory_block );
   int  terminate(  );
   int  change_priority(int new_priority, int target_process_id);
   int  request_delay( int time_delay, int wakeup_code, MsgEnv * message_envelope );
   int  send_console_chars(MsgEnv * message_envelope );
   int  get_console_chars(MsgEnv * message_envelope );
   int  get_trace_buffers( MsgEnv * message_envelope);
};
```

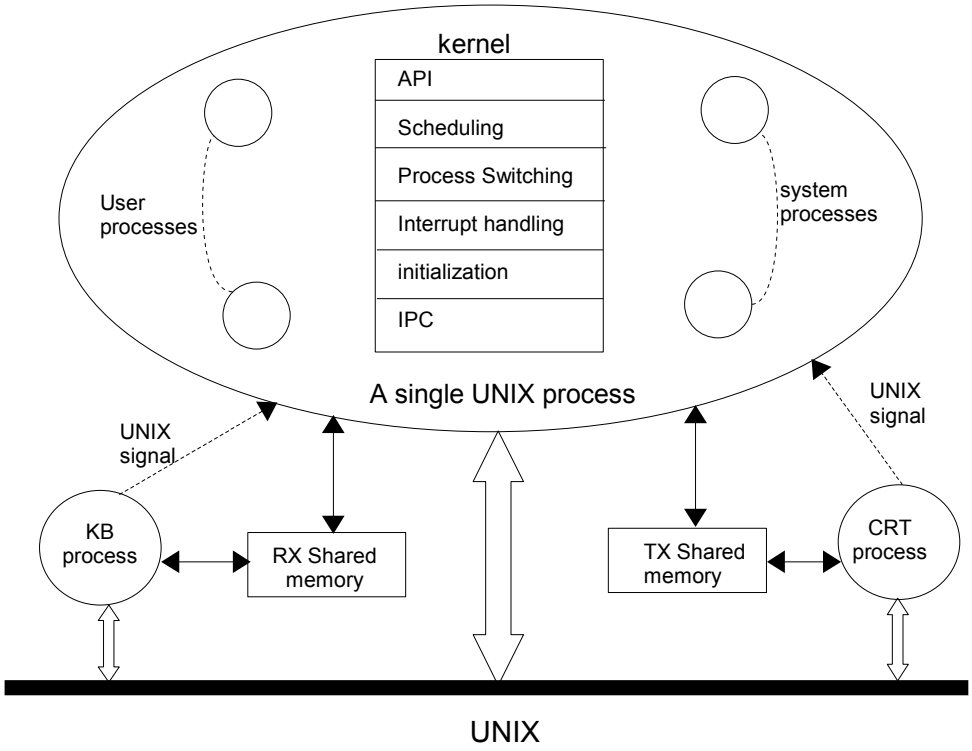
User level processes will invoke only the public methods defined in the class RTX. Thus, the public methods in class `RTX` define the user API for your real-time kernel. The MsgEnv class will also have a set of public interface methods (your design) that can be called by user processes. There is no requirement to perform some SWI to gain entry into the kernel. The user visible API methods will perform the `atomic(on/off)` functionality to be discussed shortly.

13 Appendix B: MTE 241 Project Related Constraints

See the project document for complete project requirements. The project RTX is built on top of UNIX (so we have an operating system built on top of another operating system). This provides a new mix of design related problems to consider. First, UNIX will not allow access to the interrupt handlers and not allow direct device access. As a result, the hardware UARTs must be emulated in some manner. Since there is no access to interrupts we must use some 'equivalent' paradigm. UNIX supports the concept of 'signals'. One process can signal another process, etc. Thus signals can be used to emulate device interrupts. UNIX can also provide us with a periodic signal (say every *n* ms) to emulate a timer interrupt.

13.1 The Run-Time Environment

The design must be careful as to what UNIX functionality is invoked, since UNIX might block a process on some I/O call. The real-time system (user processes + kernel) must never block on a UNIX call. So how can a user process get I/O without a UNIX call and be prevented from a UNIX block – only our kernel should be able to block a user process. Consider the following figure which shows 3 UNIX processes.



The project's user and system processes share a single UNIX process (called the RTX process) with the kernel. The KB UNIX process and the RX shared memory emulate the RX portion of a UART while the CRT UNIX process and the TX shared memory emulate the TX portion of a UART. The UART TX and RX interrupts are emulated by sending UNIX signals to the RTX process. All required I/O is directed to the KB/CRT processes allowing the RTX process to execute without being blocked on any UNIX I/O request.

The RX shared memory contains the RX UART status and the characters received from the keyboard. This shared memory can be read/written by the RTX and KB UNIX processes. Once the KB process has received one

or more characters from the keyboard, it places these in the RX shared memory and sends a signal to the RTX process. Then, the RTX process's signal handlers (i.e. interrupt handlers) can read the character(s) and start the appropriate i_process for further handling. The KB process 'polls' (in a non-busy way) the shared memory to determine when the character(s) have been removed by the RTX's signal handler. Therefore, it is the KB process that blocks on a UNIX I/O call rather than the RTX process.

The TX shared memory contains the TX UART status and the character(s) to be sent to the CRT. This shared memory can be read/written by the RTX and CRT UNIX processes. Once the CRT process has received one or more characters from the RTX process in its shared memory, it makes a possibly blocking call to UNIX to write the characters onto the crt. Once the TX shared memory has been emptied of characters, the CRT process sends a signal to the RTX process with its new status in the shared memory. Then, the RTX process's signal handlers (i.e. interrupt handlers) can put additional characters into the shared memory and so on. The CRT process 'polls' (in a non-busy way) the shared memory to determine when additional character(s) have been placed into the shared memory by the RTX process. Therefore, it is the CRT process that blocks on a UNIX I/O call rather than the RTX process.

It is important that proper synchronization methods be implemented to ensure that reading and writing of shared memory is properly coordinated. Since the KB and CRT process basically 'poll' (when they are not blocked) the contents of their shared memory, it is important that they do not just loop and examine status bits of the shared memory since they would consume most of the CPU time. For example, they could call a UNIX *sleep* function after each poll to allow the RTX process to run. This could result in a short latency from the time a status bit is set in the shared memory to the time the CRT/KB process recognizes it.

13.2 The user/kernel Interface Revisited

In the MTE 241 project, we cannot use the SWI sequence for entry into the kernel. We can define the function `atomic(on/off)` which essentially masks or unmasks the UNIX signals. Thus, `atomic(on)` would be equivalent to disabling all interrupts (signals), while `atomic(off)` would enable the signal system. This allows the concept of atomicity and allows kernel functions to execute without interference from signals. The atomic function is discussed in detail in a later section.

Consider the following generic C-language example of the usage of the atomic function. The function would be used in the user visible API functions before and after the call to the kernel version of the primitive.

```
int send_message( int dest_pid, char * msg)   // the user visible API function
{
       int return_val;

       atomic( on );
       return_val = K_send_message( dest_pid, msg);  // call the kernel version
       atomic( off );
       return return_val;
}
```

While the kernel's `K_send_message` function executes, it is protected from interference from signals.

13.3 Using the setjmp/longjmp Sequence

The UNIX setjmp and longjmp system calls are used in the project to perform process switching and to prepare a process to execute for the first time (with no previous context).

13.3.1 RTX Initialization

This section extends the information found in the on-line UNIX programming tutorial. The following is a

general overview of how to initialize the process' context during the RTX initialization. This allows the processes to run within their own context when the dispatcher selects them to execute for the *first time* (i.e. on their first dequeue from the ready process queue). The following code excerpts are not themselves compilable (they may include pseudocode, etc.); they serve as a basic guide only. You should customize to adapt to your particular design approach.

Our goal, here, is to use the initialization table (IT) to create the PCBs, initialize them, put each PCB on the process queue and the respective ready queue. During this task, we also want to create the stack for each process and prepare a context save area so the process can start "for the first time". There is a major difference in starting a process for the first time and later performing a process switch to re-start the user process.

The following operations would take place during the RTX initialization phase (i.e. all PCBs, queues, internal data structures are created and initialized).

The following example assumes there is type definition for a PCB, there is access to the initialization table and we can determine the number of user processes defined in the IT. Here, we assume that the user processes are defined first in the IT followed by any i_processes. We only initialize the user processes in the following example. The example is given in C notation and preceded by an explanation of its operation:

Comments: (for the line labels)

1. We save the kernel context temporarily into *kernel_buf* (a local variable of type jmp_buf). The set_jmp call will return with the value 0, so we proceed to line labelled 2.
2. We set the current stack pointer to have the value of the stack pointer for the process being initialized. This was set up previously in the code to initialize the PCB .
3. Here, we set up the context for the process being initialized. Since this is the first call to set_jmp on this buffer, it will return with value 0 and we proceed to statement labelled 4. But when the current process is finally chosen to run by the dispatcher (when it is dequeued from the RPQ), the long_jmp at that time will cause a 'return' to the set_jmp of statement 3 **but** its return value will be 1 and this will cause statement 5 to be executed.
4. Here, we restore the context of the kernel initialization (previously stored in *kernel_buf*). This statement will cause a return to be made to statement 1's set_jmp call **but** this time with a return value of 1 - - which causes the condition of the *if* statement to fail and then statement 6 would be the next statement executed after statement 4 causing the current iteration to complete and we start the next iteration of the outer `for` loop..
5. We come here only when the dispatcher's swap function executes a long_jmp on the target PCB's context save buffer. See the comments for statement 3. We return here only the first time the process is selected for execution. Here we assume that the variable `current_process` has been initialized (by the dispatcher) to point the the PCB of the next process to run and that its status field has already been set to '`executing`'. We get the start address of the process and then execute the process -- we should never get a return to the function call made here.

This is typical of the operations required to initialize the context of all PCBs so that their respective processes will execute properly on the first time they are removed from the RPQ. On subsequent selection by the dispatcher, the process will be 'controlled' by the set_jmp / long_jmp instructions within the swap( ) routine which is part of the process_switch sequence. See the additional material on-line (the section on set_jmp/long_jmp).


```
//=============================================

// Do the initialization of all user processes via the IT structures.
// Put all PCBs on ready queue and the process list.
// Set all the stacks and jmp buffers for a first time run of process.
// We do this ONLY for the user processes NOT the i_processes!!!!
// The array "itable" is the initialization table (IT).
// There is previous definition of a PCB as a C struct definition.

void iniz_processes ()
{
        int i, num_procs, size;
        struct PCB *apcb;
        jmp_buf kernel_buf;   // jump buffer used by kernel initialize

        iniz_rpq ();          //initialize the RPQ to empty
        num_procs = ?????????????; // set to number of user processes to initialize

        // now set up the PC's, stacks and jmp buffers of all user processes

        for (i = 0; i < num_procs; i++)
        {
                apcb = (struct PCB *) malloc (sizeof (struct PCB));//create the PCB
                apcb->priority = itable[i].priority;  // initialize fields of PCB
                apcb->proc_pid = itable[i].id;
                size = itable[i].stack_size;
                apcb->proc_stack = ((char *) malloc(size))+ size - STK_OFFSET; //stacks grow down
                apcb->start_PC = itable[i].start_PC;
                apcb->status = READY;
                apcb->msgHead = NULL;
                apcb->msgTail = NULL;
                enq_proc_list(apcb);  //put it on process list
                rpq_enque(apcb);      // put it on ready-process-queue
                                      // now set up the process context and stack
                if (setjmp (kernel_buf) == 0)     //  line 1
                {
                        jmpsp = apcb->proc_stack;
#ifdef i386
                        __asm__ ("movl %0,%%esp" :"=m" (jmpsp)); // if Linux i386 target
#endif                                            //  line 2
#ifdef __sparc
                        _set_sp( jmpsp );                    // if Sparc target (eceunix)
#endif
                        if (setjmp (apcb->context) == 0)       //  line 3
                        {
                                longjmp (kernel_buf, 1);           //  line 4
                        }
                        else
                        {
                                void (*tmp_fn) ();                 //  line 5
                                tmp_fn = (void *) current_process->start_PC;
                                tmp_fn ();     // process starts for the first time here
                        }
                }               //  line 6
        }
}
```


13.4 Performing a Process/Context Switch within UNIX

Process/context switch was discussed in section 9. Again, we take liberties in mixing C-language statements with pseudo code. This results in expressing the desired functionality but does not produce a compilable C language function. We now look at some details that can be of use in a UNIX environment. Consider again the previous process switch functionality:

```
process_switch( )
{ PCB_ptr *next_pcb;          //local variable to hold ptr to next pcb
  next_pcb = rpq_dequeue( );  //get ptr to highest priority ready process
 context_switch( current_process, next_pcb );
}
```

We can expand this function with a few more details that must be done prior to a context switch as:

```
process_switch( )
{ PCB_ptr *next_pcb, *old_pcb;    //local variable to hold ptr to next pcb
  next_pcb = rpq_dequeue( );      //get ptr to highest priority ready process
  next_pcb->state = running;      //set state field of 'next' to running
  old_pcb = current_process;
  current_process = next_pcb;     //make the next_pcb the current process
  context_switch( old_pcb->jmp_buffer, next_pcb->jmp_buffer );
}
```

It is assumed that the field '`jmp_buffer`' contains a pointer to the process's jump buffer. Recall that prior to calling process_switch, the current_process PCB has already been placed on some queue (if appropriate) and its state field has already been updated. Hence, process_switch gets the next runnable process, sets its state field to 'running' and sets its PCB as the current_process PCB before calling context_switch. At this point, we are still executing in current_process's context. Context_switch will perform the actual switch to the selected process after the current context is saved and the next context restored.

The context switch functionality is shown below:

```
// context_switch() - performs the context switch between two user processes
// Once we call longjmp(), we resume executing the "next" process
// at the point where its context was last saved with a setjmp(). */

void context_switch(jmp_buf * previous, jmp_buf * next) {
     return_code = setjmp(*previous);   // line 1
    if (return_code == 0)               // line 2
          {
                longjmp(*next,1);       // line 3
           }
 }
```

The call to *setjmp* (on the jump buffer of previous, line 1) saves the current context and returns a '0' value. This causes the l*ongjmp* to be executed on the jump buffer of next (line 3). At this point, the context of next is restored to the last point it called setjmp (which was line 1 the last time a context switch was performed). But this time a '1' value will be returned so the '`if`' condition fails and the next process executes the return from context_switch and returns to who ever called context_switch. This overview assumes that the 'next' process has executed before and was process switched in this manner. If this was the first execution of process 'next', then its *longjmp* execution (line 3) would take it back to the initialization section code where its original *setjmp* was executed.

13.5 Forking the KB and CRT Processes

Since the KB and CRT processes are separate from the RTX process, they must be forked during the RTX initialization. Typically, this would be done last after all the user and system processes have been created and the

signal handlers and signal masks have been set up. Some basic requirements during this portion of initialization is:

1. Obtain both shared memory segments from UNIX first and attach the RTX process to the shared memories.
2. The RTX process must know its UNIX process id.
3. Fork each process (KB and CRT).
4. Each forked process excel's its own program code and passes to it the UNIX process id of the RTX parent process. All excel'd processes use the appropriate memory key (the one that the RTX process used) to obtain and attach themselves to the shared memory. Then the excel'd processes can send signals to the RTX process and can access their respective shared memory.

For a detailed example, see the 'demo.c' sample later in this document.

13.5.1 POSIX Style Shared Memory

In previous years, we have used the "conventional" UNIX style of creating shared memory segments. Traditionally, this has functioned fine. Recently, most Linux and even Solaris kernels give errors when creating shared memory segments using the standard functions *shmget*, *shmat*, etc. unless the user is logged in as 'root'. This was investigated on eceunix by lab personnel with no apparent acceptable solution.

There is another way of obtaining shared memory segments by employing a subset of the POSIX standard functions. Since the POSIX standard is to ensure portability across different kernel (and OS) implementations, this should be a safe and reliable way of creating shared memory segments. The basic approach is: i) to create a temporary file for read/write access with permissions restricted to owner 'rwx' access only, ii) make the size of the file the same size as the required shared memory segment, iii) make the file memory mapped into a memory buffer the size of the required shared memory segment, and iv) obtain a pointer to the memory buffer [the pointer is now pointer to a real shared memory segment which child processes can access as well – after the child process uses the file identifier obtained from the parent process]. When the shared memory segment is no longer required, then: i) unmap the memory buffer, ii) close the temporary file and iii) unlink (i.e. delete) the temporary file. See the fragments, in a later section (from demo.c from the course web site), which illustrate the basic method. The functions of interest are: *ftruncate*, *mmap*, *munmap* and *unlink*.

Aside:

> The POSIX approach was successfully compiled and executed (with no source level changes) on the following systems: i) Debian Linux with 2.6 kernel, 2) eceunix with Solaris, and 3) a Windows XP PC using cygwin with the *cygserver* installed [as of June 2007].

13.6 Using UNIX Signals

Since there is no access to interrupts and interrupt handlers, we use UNIX signals and provide signal handlers to act as interrupt handlers. UNIX provides the facility to mask and unmask certain signals; an analogy to disabling and enabling interrupts. Hence the `atomic(on/off)` function will be used to mask and unmask interrupts on entry/exit from the kernel providing the atomicity requirements for kernel primitive execution. Since the requirements are for a non-preemptive RTX, we not need face the problem of performing a user process switch at the end of a signal handler's execution.

13.6.1 The atomic(..) function

The atomic function is used (see definition on page 33) as an example to stress the requirement for atomicity but with the UNIX constraints, however, it can be implemented and actually used to advantage since there is no direct access to the interrupts and interrupt handlers within a UNIX environment. For example, consider the following:

```
atomic(bool on) {
static sigset_t oldmask;
sigset_t newmask;
  if (on) {
           sigemptyset(&newmask);
           sigaddset(&newmask, SIGALRM); //the alarm signal
           sigaddset(&newmask, SIGINT); // the CNTRL-C
           sigaddset(&newmask, SIGUSR1);  // the CRT signal
           sigaddset(&newmask, SIGUSR2);  // the KB signal
           sigprocmask(SIG_BLOCK, &newmask, &oldmask);
     }
  } else {
       current_process->atomic_count = current_process->atomic_count - 1;
       if (current_process->atomic_count == 0) {
                 //unblock the signals
          sigprocmask(SIG_SETMASK, &oldmask, NULL);
       }
  }
}
```

For `atomic(on)`, the current signal mask is saved and a new signal mask is constructed which disables all signals of interest. Thus, the kernel primitives can execute indivisibly with no interference from any signal handlers. For `atomic(off)`, the previous signal mask is restored and all signals are enabled allowing the signal handlers to execute as needed.

13.6.2 The Signal Handlers

The signal handlers can each be a simple C-function. Since the signal handlers can only be invoked by UNIX when when they are enabled (see atomic above), they will not interfere with any kernel primitive's execution. The signal handler is the equivalent of the i_process and should be provided with a valid PCB when it executes. During initialization, UNIX is informed of the signal handler to invoke for each specified signal type. For example, consider the following typical code sequence to initialize the signal handling system. Assume there is some signal handler “die(..)” which performs an orderly shut down of the RTX system for a UNIX related abort signal. While setting the handlers, ensure that the signals are masked first. Once all the handlers have been installed and RTX initialization is complete, the signal handlers can be unmasked just before the first user process executes.

```
sigset(SIGINT,die);          // catch kill signals
sigset(SIGBUS,die);          // catch bus errors
sigset(SIGHUP,die);
sigset(SIGILL,die);          // illegal instruction
sigset(SIGQUIT,die);
sigset(SIGABRT,die);
sigset(SIGTERM,die);
sigset(SIGSEGV,die);  // catch segmentation faults
sigset(SIGUSR2,kbd_handler); // kbd_handler(.. ) is the handler for the KB signal
sigset(SIGUSR1,crt_handler); // crt_handler( ..) is the handler for the KB signal
sigset(SIGALRM,tick_handler); // tick_handler(.. ) is the handler for the clock signal
```

It is also possible to have a common signal handler which initially determines the origin of the signal and then

calls the specialized signal handler (i_process). The following gives a skeleton of a signal handler that is called on every signal type. First it determines the origin of the signal and then invokes the correct signal handler (i_process) to handle the event.

```
void exception_handler_process(int result)
{
// any pre-handling code
  pcb* temp;
  pcb* savePCB = RTX.current_process;

  switch(result) {
    case SIGINT: die();
                 break;
    case SIGALRM: RTX.current_process = RTX.pid_to_PCBptr(TIMER_I_PROC);
                  tick_handler( );
                  break;
    case SIGUSR1: //some code
                  break;
    case SIGUSR2: // some code
                  break;
    default:      printf("Unknown signal!\n");
                  die();
                  exit(1);
                  break;
  }
  RTX.current_process = savePCB;
       // any post handling code
  return;
}
```

13.7 Example: demo program

The following is a copy of the demo program found on the course web site. This program gives an example of how to fork a child process, pass arguments to the child, create and access a shared memory segment (accessible by parent and child processes), use the curses library, initialize and use the UNIX signalling facility to emulate interrupts. The course web site also has the makefile used to compile/link the sources. This demo was created and executed on a Linux system (Debian Linux, with 2.6 kernel). The sources were copied to eceunix (running Solaris) and were compiled and executed with no source changes.

++++++++++++++++++++the demo.c ++

```
// demo.cc
//
// demo program for MTE 241
// demonstrates       use of signals, curses and setting up shared memory
//+++++++++++++++++++++++++++++++++++++++++++++++++++++++++
/*  modified June 5, 2007 by P. Dasiewicz

       The example was modified to remove the traditional UNIX approach of using
       shared memory get/attach since this was suddenly causing problems on eceunix.

       It has been modified to use the POSIX-style using functions mmap etc. to
       defined a memory mapped file which can then be used as a shared memory
       segment (between multiple processes) and accessed by standard C-type pointers.
       Basically, behaves exactly the same as the traditional UNIX shared memory.


*/
//+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#include <stdlib.h>
#include <signal.h>
```



```
#include <unistd.h>                  // getpid() definition
#include <curses.h>
#include "demo.h"
#include <fcntl.h>

#include <sys/mman.h>
#include <sys/wait.h>
#include <stdio.h>



// globals
inputbuf  * in_mem_p;         // pointer to structure that is the shared memory
int     in_pid;                       // pid of keyboard child process
caddr_t mmap_ptr;
int bufsize = BUFFERSIZE;
int fid, status;              //used to create the shared memory

//+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
char * sfilename = "junkDemo";  //the name of the shared_memory file
//+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

//*********************************************************************
// routine to clean up things before terminating main program
// This stuff must be cleaned up or we have child processes and shared
//      memory hanging around after the main process terminates

void cleanup()
{

        endwin();             // close "curses" window

        // terminate child process(es)
        kill(in_pid,SIGINT);

        // remove shared memory segment and do some standard error checks
        status = munmap(mmap_ptr, bufsize);
    if (status == -1){
      printf("Bad munmap during cleanup\n");
    }
        // close the temporary mmap file
    status = close(fid);
    if (status == -1){
      printf("Bad close of temporary mmap file during cleanup\n");
    };
        // unlink (i.e. delete) the temporary mmap file
    status = unlink(sfilename);
    if (status == -1){
      printf("Bad unlink during claeanup.\n");
    }
}


//*********************************************************************
// routine to call before exitting
// This routine gets called when certain signals occur

void die(int signal)
{
        cleanup();
        printf( "\n\nSignal Received.   Leaving demo ...\n" );
        exit(0);
}

// kbd_handler
// Called by signal SIGUSR1 from keyboard reader process


void kbd_handler(int signum)
{
        inputbuf command;
```



```
        // copy input buffer
        if (in_mem_p->indata[0] != '\0')
        {
            strcpy(command.indata,in_mem_p->indata);

            // we should parse the input string and execute the command given,
            //  but for now we just echo the input
            // printw is the curses equivalent of printf
            printw("Keyboard input is: %s\n",command.indata);
            refresh();
        }

}

//***************************************************************************
int main()
{


        // catch signals so we can clean up everything before exitting
        // signals defined in /usr/include/signal.h
        // e.g. when we recieved an interrupt signal SIGINT, call die()
        sigset(SIGINT,die);           // catch kill signals
        sigset(SIGBUS,die);           // catch bus errors
        sigset(SIGHUP,die);
        sigset(SIGILL,die);           // illegal instruction
        sigset(SIGQUIT,die);
        sigset(SIGABRT,die);
        sigset(SIGTERM,die);
        sigset(SIGSEGV,die);  // catch segmentation faults



        // signal from keyboard reader is SIGUSR1 (user-defined signal)
        // When there is input from the keyboard, call the kbd_handler() routine
        signal(SIGUSR1,kbd_handler);


  /* Create a new mmap file for read/write access with permissions restricted
     to owner rwx access only */
  fid = open(sfilename, O_RDWR | O_CREAT | O_EXCL, (mode_t) 0755 );
  if (fid < 0){
    printf("Bad Open of mmap file <%s>\n", sfilename);
       exit(0);
  };

  // make the file the same size as the buffer
   status = ftruncate(fid, bufsize );
   if (status){
      printf("Failed to ftruncate the file <%s>, status = %d\n", sfilename, status );
      exit(0);
  }


        // pass parent's process id and the file id to child
        char childarg1[20], childarg2[20]; // arguments to pass to child process(es)
        int mypid = getpid();               // get current process pid

        sprintf(childarg1, "%d", mypid); // convert to string to pass to child
    sprintf(childarg2, "%d", fid);   // convert the file identifier


        // create the keyboard reader process
        // fork() creates a second process identical to the current process,
        // except that the "parent" process has in_pid = new process's ID,
        // while the new (child) process has in_pid = 0.
        // After fork(), we do execl() to start the actual child program.
        // (see the fork and execl man pages for more info)

        in_pid = fork();
        if (in_pid == 0)      // is this the child process ?
        {
```



```
            execl("./keyboard", "keyboard", childarg1, childarg2, (char *)0);
            // should never reach here
            fprintf(stderr,"demo: can't exec keyboard, errno %d\n",errno);
            cleanup();
            exit(1);
        };
        // the parent process continues executing here

        // sleep for a second to give the child process time to start
        sleep(1);

        // allocate a shared memory region using mmap
        // the child process also uses this region

    mmap_ptr = mmap((caddr_t) 0,   /* Memory location, 0 lets O/S choose */
                bufsize,              /* How many bytes to mmap */
                PROT_READ | PROT_WRITE, /* Read and write permissions */
                MAP_SHARED,    /* Accessible by another process */
                fid,           /* the file associated with mmap */
                (off_t) 0);    /* Offset within a page frame */
    if (mmap_ptr == MAP_FAILED){
      printf("Parent's memory map has failed, about to quit!\n");
         die(0);  // do cleanup and terminate
    };

        in_mem_p = (inputbuf *) mmap_ptr;   // pointer to shared memory
          // we can now use 'in_mem_p' as a standard C pointer to access
          // the created shared memory segment


        // initialize the screen to do output using the curses routines
        initscr();
        noecho();
        cbreak();      //do not wait for CR on input
        idlok(stdscr,TRUE);                  // enable scrolling on the standard screen
        scrollok(stdscr,TRUE);        // enable scrolling on the standard screen

        // print a message at a specific location
        // mvaddstr writes to the screen buffer
        // refresh writes what is in the screen buffer to the physical screen
        int row = 1;
        int col = 0;
        mvaddstr(row,col+10,"Demo program");
        mvaddstr(row+2,col,"Press any key to activate the keyboard handler.");
        mvaddstr(row+4,col,"Press CTRL-C to stop the program");
        move(row+6,col);      // position cursor for future output
        refresh();

        // now start doing whatever work you are supposed to do
        // in this case, do nothing; only the keyboard handler will do work
        while (1);

        // should never reach here, but in case we do, clean up after ourselves
        cleanup();
        exit(1);
} // main
```

++++++++++++ the Keyboard.c child ++

```
// keyboard.c
//
// Keyboard Reader - emulate a hardware interrupt
// read the keyboard and signal the parent process when a key is received
//
//++++++++++++++++++++++++
// modifed to use the POSIX-style of obtaining shared memory
// by P. Dasiewicz, June 5, 2007
//++++++++++++++++++++++++++
```



```
#include <stdio.h>
#include <signal.h>
#include "demo.h"
#include <fcntl.h>

#include <sys/mman.h>
#include <sys/wait.h>

int bufsize = BUFFERSIZE;

// do any necessary cleanup before exitting
// ( in this case, there is no cleanup to do)
// Basically, we rely on the parent process to cleanup shared memory
void in_die(int signal)
{
       exit(0);
}


int main (int argc, char * argv[])
{
       int parent_pid, fid;

       caddr_t mmap_ptr;
       inputbuf * in_mem_p;
       char c;


       // if parent tells us to terminate, then clean up first
       sigset(SIGINT,in_die);

       // get id of process to signal when we have input
       // and the file id of the memory mapped file
       // i.e. process input arguments
       sscanf(argv[1], "%d", &parent_pid );
       sscanf(argv[2], "%d", &fid );  // get the file id

       // attach to shared memory so we can pass input to
       // keyboard interrupt handler

       mmap_ptr = mmap((caddr_t) 0,   /* Memory Location, 0 lets O/S choose */
                   bufsize,/* How many bytes to mmap */
                   PROT_READ | PROT_WRITE, /* Read and write permissions */
                   MAP_SHARED,    /* Accessible by another process */
                   fid,           /* which file is associated with mmap */
                   (off_t) 0);    /* Offset in page frame */
    if (mmap_ptr == MAP_FAILED){
      printf("Child memory map has failed, KB is aborting!\n");
         in_die(0);
    }

       in_mem_p = (inputbuf *) mmap_ptr; // now we have a shared memory pointer

       // read keyboard
       do
       {
              c = getchar();

              if (c != '\n') // i.e. not carriage return
                     in_mem_p->indata[0] = c;
              else
                     in_mem_p->indata[0] = '\0';
              // "kill" sends a signal to the process with the specified pid
              kill(parent_pid,SIGUSR1);
       }
       while(1);

} // keyboard
```


++++++++++++++++++++++++the demo.h header file++++++++++++++++++++++++++++++++

```
// demo.h

#include <sys/types.h>
#include <sys/ipc.h>
#include <sys/shm.h>
#include <errno.h>

#include <string.h>
#include <stdlib.h>

#define BUFFERSIZE 100;

// structures for shared memory
#define MAXCHAR 80

typedef struct {
        char indata[MAXCHAR];
        int length;
} inputbuf;
```

14 Index

To be done.