Red Hat Enterprise Linux for Real Time 8

Understanding RHEL for Real Time

An introduction to fundamental concepts for RHEL for Real Time

Last Updated: 2022-05-17

Red Hat Enterprise Linux for Real Time 8 Understanding RHEL for Real Time

An introduction to fundamental concepts for RHEL for Real Time

Legal Notice

Abstract

This guide helps users understand the fundamental concepts and associated references for optimizing the RHEL for Real Time kernel to maintain low latency, consistent response time, and determinism.

Table of Contents

PROVIDING FEEDBACK ON RED HAT DOCUMENTATION

We appreciate your input on our documentation. Please let us know how we could make it better.

- For simple comments on specific passages:
  1. Make sure you are viewing the documentation in the *Multi-page HTML* format. In addition, ensure you see the **Feedback** button in the upper right corner of the document.
  2. Use your mouse cursor to highlight the part of text that you want to comment on.
  3. Click the **Add Feedback** pop-up that appears below the highlighted text.
  4. Follow the displayed instructions.
- For submitting feedback via Bugzilla, create a new ticket:
  1. Go to the Bugzilla website.
  2. As the Component, use **Documentation**.
  3. Fill in the **Description** field with your suggestion for improvement. Include a link to the relevant part(s) of documentation.
  4. Click **Submit Bug**.

CHAPTER 1. HARDWARE PLATFORMS FOR RHEL FOR REAL TIME

Configuring hardware correctly plays a critical role in setting up the real-time environment as hardware impacts the way your system would operate. Not all hardware platforms are real-time capable and enable fine tuning. Before performing fine tuning, you must ensure that the potential hardware platform is real-time capable.

Hardware platforms might vary from vendor to vendor. You can test and verify the hardware suitability for real-time with the hardware latency detector (**hwlatdetect**) program. It controls the latency detector kernel module and helps to detect latencies caused by underlying hardware or firmware behavior.

Prerequisites

- Ensure that the RHEL-RT packages are installed.
- Check the vendor documentation for any tuning steps required for low latency operation. The vendor documentation can provide instructions to reduce or remove any System Management Interrupts (SMIs) that would transition the system into System Management Mode (SMM).

NOTE

Red Hat recommends to not completely disable System Management Interrupts (SMIs), as it can result in catastrophic hardware failures.

1.1. PROCESSOR CORES

A processor core is a physical Central Processing Unit (CPU) and the core executes the machine code. A socket is a connection between the processor and the motherboard of the computer. The socket is the location on the motherboard that the processor is placed into. There are two sets of processors: single core processor that physically occupies one socket with one available core and quad-core processor occupying one socket with four available cores.

When designing a real time environment, be aware of the number of available cores, the cache layout among cores, and how the cores are physically connected.

When multiple cores are available, use threads or processes. A program when written without using these constructs, runs on a single processor at a time. A multi-core platform provides advantages through using different cores for different types of operations.

Caches

Caches have a noticeable impact on overall processing time and determinism. Often, the threads of an application need to synchronize access to a shared resource, such as a data structure.

With the **tuna** command line tool (CLI), you can determine the cache layout and bind interacting threads to a core so that they share the cache. Cache sharing reduces memory faults by ensuring that the mutual exclusion primitive (mutex, condition variables, or similar) and the data structure use the same cache.

Interconnects

Increasing the number of cores on systems can cause conflicting demands on the interconnects. This makes it necessary to determine the interconnect topology to help detect the conflicts that occur between the cores on real-time systems.

Many hardware vendors now provide a transparent network of interconnects between cores and memory, known as Non-uniform memory access (NUMA) architecture.

NUMA is a system memory design used in multiprocessing, where the memory access time depends on the memory location relative to the processor. When you use NUMA, a processor can access its own local memory faster than non-local memory, such as memory on another processor or memory shared between processors. On NUMA systems, understanding the interconnect topology helps to place threads that communicate frequently on adjacent cores.

The **taskset** and **numactl** utilities determine the CPU topology. **taskset** defines the CPU affinity without NUMA resources such as memory nodes and **numactl** controls the NUMA policy for processes and shared memory.

1.2. ADDITIONAL RESOURCES

- Installing RHEL 8 for Real Time

CHAPTER 2. MEMORY MANAGEMENT ON RHEL FOR REAL TIME

Real time systems use a virtual memory system, where an address referenced by a user-space application translates into a physical address. This happens through a combination of page tables and address translation hardware in the underlying computer system.

An advantage of having the translation mechanism in between a program and the actual memory is that the operating system can swap pages when required or on CPU demand.

For swapping pages from secondary storage to the main memory, the previously used page table entry is marked as invalid. Hence, even under normal memory pressure, the operating system might scavenge pages from one application and give them to another. This can cause unpredictable system behaviors.

Memory allocation implementations include demand paging mechanism and memory lock system calls (**mlock()**).

NOTE

Sharing data information on CPUs in different cache and NUMA domains might cause traffic problems and bottlenecks.

When writing a multithreaded application, consider the machine topology before designing the data decomposition. Topology is the memory hierarchy, and includes CPU caches and the Non-Uniform Memory Access (NUMA) node.

2.1. DEMAND PAGING

Demand paging is similar to a paging system with page swapping. The system loads pages that are stored in the secondary memory when required or on CPU demand. All memory addresses generated by a program pass through an address translation mechanism in the processor. The addresses then convert from a process-specific virtual address to a physical memory address. This is referred to as virtual memory. The two main components in the translation mechanism are page tables and translation lookaside buffers (TLBs)

Page tables

Page tables are multi-level tables in physical memory that contain mappings for virtual to physical memory. These mappings are readable by the virtual memory translation hardware in the processor.

A page table entry with an assigned physical address, is referred to as the resident working set. When the operating system needs to free memory for other processes, it can swap pages from the resident working set. When swapping pages, any reference to a virtual address within that page creates a page fault and causes page reallocation.

When the system is extremely low on physical memory, the swap process starts to thrash, which constantly steals pages from processes, and never allows a process to complete. You can monitor the virtual memory statistics by looking for the **pgfault** value in the **/proc/vmstat** file.

Translation lookaside buffers

Translation Lookaside Buffers (TLBs) are hardware caches of virtual memory translations. Any processor core with a TLB checks the TLB in parallel with initiating a memory read of a page table entry. If the TLB entry for a virtual address is valid, the memory read is aborted and the value in the TLB is used for the address translation.

A TLB operates on the principle of locality of reference. This means that if code stays in one region of memory for a significant period of time (such as loops or call-related functions) then the TLB references avoid the main memory for address translations. This can significantly speed up processing times.

When writing deterministic and fast code, use functions that maintain locality of reference. This can mean using loops rather than recursion. If recursion are not avoidable, place the recursion call at the end of the function. This is called tail-recursion, which makes the code work in a relatively small region of memory and avoids calling table translations from the main memory.

2.2. MAJOR AND MINOR PAGE FAULTS

RHEL for Real Time allocates memory by breaking physical memory into chunks called pages and then maps them to the virtual memory. Faults occur when a process needs a specific page that is not mapped or is no longer available in the memory. Hence faults essentially mean unavailability of pages when required by a CPU. When a process encounters a page fault, all threads freeze until the kernel handles this fault. There are several ways to address this problem, but the best solution can be to adjust the source code to avoid page faults.

Minor page faults

Minor page faults occur when a process attempts to access a portion of memory before it has been initialized. In such scenarios, the system performs operations to fill the memory maps or other management structures. The severity of a minor page fault can depend on the system load and other factors, but they are usually short and have a negligible impact.

Major page faults

A severe memory latency is a major page fault. Major faults occur when the system has to either synchronize memory buffers with the disk, swap memory pages belonging to other processes, or undertake any other input output (I/O) activity to free memory. This occurs when the processor references a virtual memory address that does not have a physical page allocated to it. The reference to an empty page causes the processor to execute a fault, and instructs the kernel code to allocate a page which increases latency dramatically.

When an application shows a performance drop, it is beneficial to check for the process information for page faults in the **/proc/** directory. For a particular specific process identifier (PID), use the **cat** command to view the **/proc/PID/stat** file for following relevant entries:

- Field 2: the executable file name.
- Field 10: the number of minor page faults.
- Field 12: the number of major page faults.

The following example demonstrates viewing page faults with the **cat** command and a **pipe** function.

```
# cat /proc/3366/stat | cut -d\ -f2,10,12
(bash) 5389 0
```

In the example output, the process with PID 3366 is bash and it has 5389 minor page faults and no major page faults.

Additional resources

- Linux System Programming by Robert Love

2.3. MLOCK SYSTEM CALLS

The memory lock (**mlock**) system calls enable calling processes to lock or unlock a specified range of the address space. This prevents Linux from paging the locked memory to swap space. Once you allocate the physical page to the page table entry, references to that page is relatively fast. There are two groups of **mlock** system calls: **mlock** and **munlock** calls.

mlock and **munlock** system calls lock and unlock a specific range of process address pages. When successful, the pages in the specified range remain resident in the memory until **munlock** unlocks the pages.

mlock and **munlock** take the following parameters:

- **addr**: specifies the start of an address range.
- **len**: specifies the length of the address space in bytes.

When successful, **mlock** and **munlock** return 0. In case of an error, they return -1 and set a **errno** to indicate the error.

The **mlockall** and **munlockall** calls lock or unlock the entire program space.

NOTE

The use of **mlock** does not guarantee that the program will experience no page I/O. It ensures that the data remains in the memory but cannot ensure that it remains in the same page. Other functions such as **move_pages** and memory compactors can move data around despite the use of **mlock**.

Memory locks are made on a page basis and do not stack. It means that if two dynamically allocated memory segments share the same page locked twice by calls to **mlock** or **mlockall**, they are unlocked by a single call to **munlock** or to **munlockall**. Thus, it is important to be aware of the pages the application unlocks to prevent double-lock or single-unlock problems.

The two most common alternatives to mitigate double-lock or single-unlock problems are:

- Tracking the memory areas allocated and locked and creating a wrapper function, which before unlocking a page, verifies the number of allocations the page has. This is the resource counting principle used in device drivers.
- Performing allocations considering the page size and alignment, in order to prevent a double-lock in the same page.

Additional resources

- **capabilities(7)** man page
- **mlock(2)** man page
- **mlock(3)** man page
- **mlockall(2)** man page
- **mmap(2)** man page
- **move_pages(2)** man page

- **posix_memalign(3)** man page
- **posix_memalign(3p)** man page

2.4. SHARED LIBRARIES

Shared libraries are called dynamic shared objects (DSO) and are a collection of pre-compiled code blocks called functions. These functions are reusable in multiple programs and they load at run-time or compile time.

Linux supports the following two library classes:

- Dynamic or shared libraries: exists as separate files outside of the executable file. These files load into the memory and get mapped at run-time.
- Static libraries: are files linked to a program statically at compile time.

The **ld.so** dynamic linker loads the shared libraries required by a program and then executes the code. The DSO functions load the libraries in the memory once and multiple processes can then reference the objects by mapping into the address space of processes. You can configure the dynamic libraries to load at compile time using the **LD_BIND_NOW** variable.

Evaluating symbols before program initialization can improve performance because evaluating at application run-time can cause latency if the memory pages are located on an external disk.

Additional resources

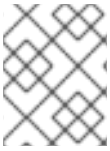
- **ld.so(8)** man page

2.5. SHARED MEMORY

Shared memory is a memory space shared between multiple processes. Using program threads, all threads created in one process context can share the same address space. This makes all data structures accessible to threads. With POSIX shared memory calls, you can configure processes to share a part of the address space.

Following are supported POSIX shared memory calls:

- **shm_open()**: creates and opens a new or opens an existing POSIX shared memory object.
- **shm_unlink**: unlinks POSIX shared memory objects.
- **mmap**: creates a new mapping in the virtual address space of the calling process.



NOTE

The mechanism for sharing a memory region between two processes using System V IPC **shmem** set of calls is deprecated and is no longer supported on RHEL for Real Time.

Additional resources

- **shm_open(3)** man page
- **shm_overview(7)** man page

- **mmap(2)** man page

CHAPTER 3. HARDWARE INTERRUPTS ON RHEL FOR REAL TIME

A standard system receives many millions of interrupts over the course of its operation, including a semi-regular "timer" interrupt that periodically performs maintenance and system scheduling decisions. It may also receive special kinds of interrupts, such as Nonmaskable Interrupt (NMI) and System Management Interrupt (SMI).

Hardware interrupts are used by devices to indicate a change in the physical state of the system that requires attention. For example, a hard disk signaling that it has read a series of data blocks, or when a network device has processed a buffer containing network packets.

Hardware interrupts are referenced by an interrupt number. These numbers are mapped back to the piece of hardware that created the interrupt. This enables the system to monitor which device created the interrupt and when it occurred.

When an interrupt occurs, active programs are stopped and an interrupt handler is executed. The handler preempts other running programs and system activities. This can slow the entire system and create latencies. RHEL for Real Time modifies the way interrupts are handled in order to improve performance and decrease latency. Using the **cat /proc/interrupts** command you can print an output to view the types of hardware interrupts that took place, the number of interrupts received, the target CPU for the interrupt, and the device generating the interrupt.

3.1. LEVEL-SIGNALED INTERRUPTS

Level-signaled interrupts use a dedicated interrupt line that delivers voltage transitions. The device controller raises an interrupt by asserting a signal on the interrupt request line. The interrupt line sends one of two voltages to represent a binary 1 or binary 0.

Once the interrupt signal is sent by the line, it remains in that state until the CPU resets it. The CPU performs a state save, captures the interrupt, and dispatches the interrupt handler. The interrupt handler determines the cause of the interrupt, clears the interrupt by performing necessary services, and restores the state of the device.

While more complex to implement, the Level-signaled interrupts are more reliable and supports multiple devices.

3.2. MESSAGE-SIGNALED INTERRUPTS

Many modern systems use message-signaled interrupts (MSI), which send the signal as a dedicated message on a packet or message-based electrical bus. One common example of this type of bus is the Peripheral Component Interconnect Express (PCI Express or PCIe). These devices transmit a message as a type that the PCIe host controller interprets as an interrupt message. The host controller then sends the message on to the CPU.

Depending on the hardware, a PCIe system might send the signal using a dedicated interrupt line between the PCIe host controller and the CPU, or send the message over the CPU HyperTransport bus. Many PCIe systems can also operate in legacy mode, where legacy interrupt lines are implemented in order to support older operating systems or on boot Linux kernels with the option **pci=nomsi** on the kernel command line.

3.3. NON-MASKABLE INTERRUPTS

Non-maskable interrupts are hardware interrupts that standard interrupt masking techniques in the system cannot ignore. NMIs have higher priority than maskable interrupts and they occur to signal attention for non recoverable hardware errors.

NMIs are also used by some systems as a hardware monitor. When a processor receives NMI, it handles the NMI immediately by calling the NMI handler pointed to by the interrupt vector. If certain conditions are met, such as an interrupt not being triggered after a specified length of time, the NMI handler signals a warning and provides debugging information about the problem. This helps to identify and prevent system lockups.

Maskable interrupts are hardware interrupts that can be ignored by setting a bit in an interrupt mask register's bit-mask. CPUs can temporarily ignore maskable interrupts during critical processing.

3.4. SYSTEM MANAGEMENT INTERRUPTS

System management interrupts (SMIs) offer extended functionality, such as legacy hardware device emulation and can also be used for system management tasks. SMIs are similar to non maskable interrupts (NMIs) in that they use a special electrical signalling line and are generally not maskable.

When an SMI occurs, the CPU enters the System Management Mode (SMM). In this mode, a special low-level handler executes to handle the SMIs. The SMM is typically provided directly from the system management firmware, often the BIOS or the EFI.

SMIs are most often used to provide legacy hardware emulation. A common example is to imitate a diskette drive. When there is no diskette drive attached, the operating system attempts to access the diskette and triggers a SMI. In this scenario, a handler provides the operating system with an emulated device instead. The operating system then treats the emulation as a legacy device.

SMIs can adversely affect the real-time systems because they take place without the direct involvement of the operating system. A poorly written SMI handling routine may consume many milliseconds of CPU time, and the operating system might not be able to preempt the handler. This can create periodic high latencies in an otherwise well-tuned and highly responsive system.

As a vendor may use SMI handlers to manage CPU temperature and fan control, it might not be possible to disable them. In such situations, you must notify the vendor of problems that occur when using these interrupts.

You can isolate SMIs on a real-time system using the **hwlatdetect** utility. It is available in the **rt-tests** package. This utility measures the time period during which the CPU is used by an SMI handling routine.

3.5. ADVANCED PROGRAMMABLE INTERRUPT CONTROLLER

The advanced programmable interrupt controller (APIC) developed by Intel Corporation, provides the ability to:

- Handle large amounts of interrupts to route each to a specific set of CPUs.
- Support inter-CPU communication and remove the need for multiple devices to share a single interrupt line.

APIC represents a series of devices and technologies that work together to generate, route, and handle a large number of hardware interrupts in a scalable and manageable way. It uses a combination of a local APIC built into each system CPU and a number of Input/Output APICs that are connected directly to hardware devices.

When a hardware device generates an interrupt, the connected I/O APIC detects it and routes across

the system APIC bus to a specific CPU. The operating system knows which IO-APIC is connected to which device, and interrupt line within that device. The Advanced Configuration and Power Interface Differentiated System Description Table (ACPI DSDT) includes information about the specific wiring of the host system motherboard and peripheral components and a device provides information on the available interrupt sources. Together, these two sets of data provide information about the overall interrupt hierarchy.

Complex APIC-based interrupt management strategies are possible, with the system APICs connected in hierarchies and delivering interrupts to CPUs in a load-balanced fashion rather than targeting a specific CPU or set of CPUs.

CHAPTER 4. RHEL FOR REAL TIME PROCESSES AND THREADS

The key factors in RHEL for Real Time operating systems are minimal interrupt latency and minimal thread switching latency. Although all programs use threads and processes, RHEL for Real Time handles them in a different way compared to the standard Red Hat Enterprise Linux.

In real-time, using parallelism helps achieve greater efficiency in task execution and latency. Parallelism is when multiple tasks or several sub-tasks run at the same time using the multi-core infrastructure of CPU.

4.1. PROCESSES

A process in real-time, in simplest terms, is a program in execution. The term process refers to an independent address space, potentially containing multiple threads. When the concept of more than one process running inside one address space was developed, Linux turned to a process structure that shares an address space with another process. This works well, as long as the process data structure is small.

A UNIX®-style process construct contains:

- Address mappings for virtual memory.
- An execution context (PC, stack, registers).
- State and accounting information.

Each process starts with a single thread, often called the parent thread. You can create additional threads from parent threads using the **fork()** system calls.

fork() creates a new child process which is identical to the parent process except for the new process identifier. The child process runs independent of the creating process. The parent and child processes can be executed simultaneously.

The difference between the **fork()** and **exec()** system calls is that, **fork()** starts a new process which is the copy of the parent process and **exec()** replaces the current process image with the new one.

When successful, **fork()** returns the process identifier of the child process and the parent process returns a non-zero value. On error, it returns an error number.

4.2. THREADS

In real-time, multiple threads can exist within a process. All threads of a process share its virtual address space and system resources. A thread is a schedulable entity that contains:

- A program counter (PC).
- A register context.
- A stack pointer.

Following are potential mechanisms to create parallelism:

- Using the **fork()** and **exec** function calls to create new processes. **fork()** creates an exact duplicate of a process from which it is called and has a unique process identifier.

- Using the Posix threads (**pthreads**) API to create new threads within an already running process.

You must evaluate the component interaction level before forking threads. Creating a new address space and running it as a new process is beneficial when the components are independent of one another or with less interaction.

When components are required to share data or communicate frequently, running them as threads within one address space is more efficient.

When successful, **fork()** returns a zero value. On error, it returns an error number.

4.3. ADDITIONAL RESOURCES

- **fork(2)** man page
- **exec(2)** man page

CHAPTER 5. APPLICATION TIMESTAMPING ON RHEL FOR REAL TIME

Applications that perform frequent **timestamps** are affected by the CPU cost of reading the clock. The high cost and amount of time used to read the clock can have a negative impact on an application's performance.

You can reduce the cost of reading the clock by selecting a hardware clock that has a reading mechanism, faster than that of the default clock.

In RHEL for Real Time, a further performance gain can be acquired by using POSIX clocks with the **clock_gettime()** function to produce clock readings with the lowest possible CPU cost.

These benefits are more evident on systems which use hardware clocks with high reading costs.

5.1. HARDWARE CLOCKS

Multiprocessor systems such as Non-uniform memory access (NUMA) or Symmetric Multi Processing (SMP) have multiple instances of clock sources. The interaction between all clock sources and their response to system events such as CPU frequency scaling or entering energy economy modes, determines the clock sources suitable for the real-time kernel.

During boot time the kernel selects a suitable clock source from the list of available clock sources that is maintained in the **/sys/devices/system/clocksource/clocksource0/available_clocksource** file. The clock source currently in use is available in the **/sys/devices/system/clocksource/clocksource0/current_clocksource** file.

The preferred clock source is the Time Stamp Counter (TSC), but if it is not available the High Precision Event Timer (HPET) is the second best option. However, not all systems have HPET clocks and some HPET clocks can be unreliable.

In the absence of TSC and HPET, other options include the ACPI Power Management Timer (ACPI_PM), the Programmable Interval Timer (PIT) and the Real Time Clock (RTC). The PIT and RTC sources can be either costly to read or have a low resolution (time granularity). Therefore they are sub-optimal for the real-time kernel.

Though TSC is generally the preferred clock source, some of its hardware implementations can have shortcomings. For example, some TSC clocks can stop when the system moves to a idle state, or become out of sync when the CPU enter deeper C-states (energy saving states) or perform speed or frequency scaling operations. You can workaround these issues by using additional kernel boot parameters, such as the **idle=poll** and **max_cstate=1** parameters. The **idle=poll** parameter forces the clock to avoid the idle state and **max_cstate=1** prevents the clock from entering deeper C states. C states are states when the CPU reduces or turns off certain functions and idle state is when a CPU is not being used by any program.

As an example program, the following output shows the list of the available clock sources in your system.

```
# cat /sys/devices/system/clocksource/clocksource0/available_clocksource tsc hpet acpi_pm
```

5.2. POSIX CLOCKS

POSIX is a standard for implementing and representing time sources. You can assign a POSIX clock to an application without affecting other applications in the system. This is in contrast to hardware clocks which are selected by the kernel and implemented across the system.

The function used to read a given POSIX clock is **clock_gettime()**, which is defined at **<time.h>**. The kernel counterpart to **clock_gettime()** is a system call. When a user process calls **clock_gettime()**:

1. The corresponding C library (**glibc**) calls the **sys_clock_gettime()** system call.
2. **sys_clock_gettime()** performs the requested operation.
3. **sys_clock_gettime()** returns the result to the user program.

However, the context switch from the user application to the kernel has a CPU cost. Even though this cost is very low, if the operation is repeated thousands of times, the accumulated cost can have an impact on the overall performance of the application. To avoid context switching to the kernel, thus making it faster to read the clock, support for the **CLOCK_MONOTONIC_COARSE** and **CLOCK_REALTIME_COARSE** POSIX clocks was added, in the form of a virtual dynamic shared object (VDSO) library function.

Time readings performed by **clock_gettime()**, using one of the **_COARSE** clock variants, do not require kernel intervention and are executed entirely in user space. This yields a significant performance gain. Time readings for **_COARSE** clocks have a millisecond (ms) resolution, meaning that time intervals smaller than 1 ms are not recorded. The **_COARSE** variants of the POSIX clocks are suitable for any application that can accommodate millisecond clock resolution.

5.3. CLOCK_GETTIME FUNCTION

The following code shows an example of code using the **clock_gettime** function with the **CLOCK_MONOTONIC_COARSE** POSIX clock:

```
#include <time.h>

main()
{
 int rc;
 long i;
 struct timespec ts;

 for(i=0; i<10000000; i++) {
  rc = clock_gettime(CLOCK_MONOTONIC_COARSE, &ts);
 }
}
```

You can improve upon the example above by adding checks to verify the return code of **clock_gettime()**, to verify the value of the **rc** variable, or to ensure the content of the **ts** structure is to be trusted.

NOTE

The **clock_gettime()** man page provides more information on writing more reliable applications.

IMPORTANT

Programs using the **clock_gettime()** function must be linked with the **rt** library by adding **-lrt** to the **gcc** command line.

$ gcc clock_timing.c -o clock_timing -lrt

5.4. ADDITIONAL RESOURCES

- **clock_gettime()** man page
- Linux System Programming by Robert Love
- Understanding The Linux Kernel by Daniel P. Bovet and Marco Cesati

CHAPTER 6. SCHEDULING POLICIES FOR RHEL FOR REAL TIME

The scheduler is the kernel component that determines the runnable thread to execute. Each thread has an associated scheduling policy and a static scheduling priority (**sched_priority**).

The scheduling being preemptive, the currently running thread stops when a thread with a higher static priority gets ready to execute. The current thread then returns to the **waitlist** for its static priority.

All Linux threads have one of the following scheduling policies:

- **SCHED_OTHER** or `SCHED_NORMAL: the default policy
- **SCHED_BATCH**: similar to **SCHED_OTHER**, but with a throughput orientation
- **SCHED_IDLE**: a lower priority policy than **SCHED_OTHER**.
- **SCHED_FIFO**: a first in and first out real-time policy.
- **SCHED_RR**: a round-robin real-time policy.
- **SCHED_DEADLINE**: a scheduler policy to prioritize tasks according to it's job deadline, the earliest absolute deadline runs first.

6.1. SCHEDULER POLICIES

The real-time threads have higher priority than the normal threads. The policies have scheduling priority values that range from the minimum value of 1 to the maximum value of 99.

The following policies are critical to real-time:

- **SCHED_OTHER** or **SCHED_NORMAL**:
  This is the default scheduling policy for Linux threads. It has a dynamic priority that is changed by the system based on the characteristics of the thread. **SCHED_OTHER** threads have nice values between 20 (highest priority) and 19 (lowest priority). By default, **SCHED_OTHER** threads have a nice value of 0.

- **SCHED_FIFO** policy
  Threads with **SCHED_FIFO**, run ahead of **SCHED_OTHER** tasks. Instead of using nice values, **SCHED_FIFO** uses a fixed priority between 1 (lowest) and 99 (highest). A **SCHED_FIFO** thread with a priority of 1 always schedules ahead of any **SCHED_OTHER** thread.

- **SCHED_RR** policy:
  The **SCHED_RR** policy is similar to the **SCHED_FIFO** policy. The threads of equal priority are scheduled in a round-robin fashion. **SCHED_FIFO** and **SCHED_RR** threads run until one of the following events occurs:

  - The thread goes to sleep or waits for an event.
  - A higher-priority real-time thread gets ready to run.
    Unless one of the above events occur, the threads run indefinitely on the specified processor, while the lower-priority threads remain in the queue waiting to run. This can cause the system service threads to be resident and prevent being swapped out and fail the filesystem data flushing.

- **SCHED_DEADLINE**: **SCHED_DEADLINE** policy specifies the timing requirements. It schedules each task according to the task's deadline, the task with the earliest deadline runs first. The kernel requires **runtime⇐deadline⇐period** to be true. The relation between the required options is **runtime⇐deadline⇐period**.

6.2. PARAMETERS FOR SCHED_DEADLINE POLICY

Each **SCHED_DEADLINE** task is characterized by **period**, **runtime**, and **deadline** parameters. The values for these parameters are integers of nanoseconds. The following table lists these parameters.

Table 6.1. SCHED_DEADLINE parameters

Parameter	Description
period	The activation pattern of the real-time task. For example, if a video processing task has 60 frames per second to process, a new frame will be queued for service every 16 milliseconds, so the period is 16 milliseconds.
runtime	The amount of CPU execution time allotted to the task to produce output. For a typical real-time task, the maximum execution time, also known as "Worst Case Execution Time" (WCET) is the run time. For example, a video processing tool may take, in the worst case, five milliseconds to process an image. Hence its run time is five milliseconds.
deadline	The maximum time for the output to be produced. For example, if the task needs to deliver the processed frame within ten milliseconds, the deadline will be ten milliseconds.

CHAPTER 7. AFFINITY IN RHEL FOR REAL TIME

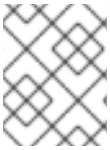
Every thread and interrupt source in the system has a processor affinity property. The operating system scheduler uses this information to determine which threads and interrupts to run on which CPU.

Affinity is represented as a bitmask, where each bit in the mask represents a CPU core. If the bit is set to 1, then the thread or interrupt may run on that core; if 0 then the thread or interrupt is excluded from running on the core. The default value for an affinity bitmask is all ones, meaning the thread or interrupt may run on any core in the system.

By default, processes can run on any CPU. However, processes can be instructed to run on a predetermined selection of CPUs, by changing the affinity of the process. Child processes inherit the CPU affinities of their parents.

Some of the more typical affinity setups include:

- Reserve one CPU core for all system processes and allow the application to run on the remainder of the cores.
- Allow a thread application and a given kernel thread (such as the network **softirq** or a driver thread) on the same CPU.
- Pair producer and consumer threads on each CPU.



NOTE

It is recommended that affinity settings must be designed in conjunction with the program, to better match the expected behavior.

7.1. PROCESSOR AFFINITY

The processes by default, can run on any CPU. However, you can configure the processes to run on a predetermined selection of CPUs, by changing the affinity of the process. Child processes inherit the CPU affinities of their parents.

The usual practice for tuning affinities on a real-time system is to determine the number of cores required to run the application and then isolating those cores. This can be achieved with the Tuna tool, or with shell scripts to modify the bitmask value.

The **taskset** command can be used to change the affinity of a process and modifying the **/proc/** filesystem entry changes the affinity of an interrupt. Using the **taskset** command with the **-p** or **--pid** option and the process identifier (PID) of the process, checks the affinity of a process.

The **-c** or **--cpu-list** option displays the numerical list of cores, instead of as a bitmask. The affinity can be set by specifying the number of the CPU to bind a specific process. For example, for a process that previously used either CPU 0 or CPU 1, you can change the affinity so that it can only run on CPU 1.

In addition to the **taskset** command, you can also set the processor affinity can using the **sched_setaffinity()** system call.

Additional resources

- **taskset(1)** man page
- **sched_setaffinity(2)** man page

7.2. SCHED_DEADLINE AND CPUSETS

SCHED_DEADLINE implements early deadline first scheduler (EDF) for sporadic tasks with a constrained deadline. It prioritizes the tasks according to the task's job deadline: earliest absolute deadline first. In addition to the EDF scheduler, the deadline scheduler also implements the constant bandwidth server (CBS). The CBS algorithm is a resource reservation protocol.

The CBS guarantees that each task receives its run time **(Q)** at every period **(T)**. At the start of every activation of a task, the CBS replenishes the task's run time. As the job runs, it consumes its **runtime** and if the task runs out of its **runtime**, the task is throttled and de-scheduled. The throttling mechanism prevents a single task from running more than its runtime and helps to avoid the performance problems of other jobs.

To avoid the overloading the system with **deadline** tasks, the **deadline** scheduler implements an acceptance test, which is run every time a task is configured to run with the **deadline scheduler**. The acceptance test guarantees that **SCHED_DEADLINE** tasks does not use more CPU time than the specified on the **kernel.sched_rt_runtime_us**/**kernel.sched_rt_period_us** files, which is 950 ms over 1s, by default.

CHAPTER 8. THREAD SYNCHRONIZATION MECHANISMS IN RHEL FOR REAL TIME

When two or more threads need access to a shared resource at the same time, the threads coordinate using the thread synchronization mechanism. Thread synchronization ensures that only one thread uses the shared resource at a time. The three thread synchronization mechanisms used on Linux: Mutexes, Barriers, and Condition variables (**condvars**).

8.1. MUTEXES

Mutex derives from the terms mutual exclusion. The mutual exclusion object synchronizes access to a resource. It is a mechanism that ensures only one thread can acquire a mutex at a time.

The **mutex** algorithm would serialize access to each section of code, so that only one thread of an application is running the code at any one time. Mutexes are created with the help of an attribute object known as the **mutex** attribute object. It is an abstract object, which contains several attributes that depends on the POSIX options you choose to implement. The attribute object is defined with the **pthread_mutex_t** variable. The object stores the attributes defined for the mutex.

When successful, the **pthread_mutex_init(&my_mutex, &my_mutex_attr)**, **pthread_mutexattr_setrobust()** and **pthread_mutexattr_getrobust()** functions return 0. On error, they return the error number.

After creating mutexes, you can either retain the attribute object to initialize more mutexes of the same type or you can clean up (destroy) the attribute object. The mutex is not affected in either case.

Mutexes include the standard and advanced type of mutexes.

Standard mutexes

Standard mutexes are private, non-recursive, non-robust, and non-priority inheritance capable mutexes. Initializing a **pthread_mutex_t** using **pthread_mutex_init(&my_mutex, &my_mutex_attr)** creates a standard mutex. When using the standard mutex type, your application may not benefit from the advantages provided by the **pthreads** API and the RHEL for Real Time kernel.

Advanced mutexes

Mutexes defined with additional capabilities are called advanced mutexes. Advanced capabilities include priority inheritance, robust behavior of a mutex, and shared and private mutexes. For example, for robust mutexes, initializing the **pthread_mutexattr_setrobust()** function, sets the robust attribute.

Similarly, using the attribute **PTHREAD_PROCESS_SHARED**, allows any thread to operate on the mutex, provided the thread has access to its allocated memory. The attribute **PTHREAD_PROCESS_PRIVATE** sets a private mutex.

A non-robust mutex does not release automatically and stays locked until you manually release it.

Additional resources

- **futex(7)** man page
- **pthread_mutex_destroy(P)** man page

8.2. BARRIERS

Barriers operate in a very different way when compared to other thread synchronization methods. Instead of serializing access to code regions, barriers define a point in the code where all active threads stop until all threads and processes reach this barrier. Barriers are used in situations when a running application needs to ensure that all threads have completed specific tasks before execution can continue.

Barriers take two variables. The first variable records the **stop** and **pass** state of the barrier. The second variable records the total number of threads that enter the barrier. The barrier sets the state to **pass** only when the specified number of threads reach the defined barrier. When the barrier state is set to **pass**, the threads and processes proceed further.

The **pthread_barrier_init()** function allocates the required resources to use the defined barrier and initializes it with the attributes referenced by the **attr** attribute object .

When successful, the **pthread_barrier_init()** and **pthread_barrier_destroy()** functions return the zero value. On error, it returns the error number.

8.3. CONDITION VARIABLES

Condition variables (**condvar**) is a POSIX thread construct that waits for a particular condition to be achieved before proceeding. In general, the signaled condition relates to the state of data that the thread shares with another thread. For example, a **condvar** can be used to signal that a data entry has been put into a processing queue and a thread waiting to process data from the queue can now proceed. Using **pthread_cond_init()** initializes a condition variable.

When successful, the **pthread_cond_init()**, **pthread_cond_wait()**, and **pthread_cond_signal()** functions return the zero value. On error, it returns the error number.

8.4. MUTEX CLASSES

The following table lists the mutex classes to consider when writing or porting an application.

Table 8.1. Mutex options

Advanced mutexes	Description
Shared mutexes	Defines shared access for multiple threads to acquire a mutex at a given time. Shared mutexes can create a high additional overhead. The attribute is **PTHREAD_PROCESS_SHARED**.
Private mutexes	Ensures that only the threads created within the same process can access the mutex. The attribute is **PTHREAD_PROCESS_PRIVATE**.
Real-time priority inheritance	Sets the priority level of the lower priority task higher above a current higher priority task. When the task completes, it releases the resource and the task drops back to its original priority permitting the higher priority task to execute. The attribute is **PTHREAD_PRIO_INHERIT**.

Advanced mutexes	Description
Robust mutexes	Sets the robust mutexes to release automatically when the owning thread terminates. The value substring **NP** in the string **PTHREAD_MUTEX_ROBUST_NP**, indicates that robust mutexes are non-POSIX or not portable

Additional resources

- **futex(7)** man page

8.5. THREAD SYNCHRONIZATION FUNCTIONS

The following tables lists the functions applicable for thread synchronization mechanisms

Table 8.2. Functions

Function	Description
pthread_mutexattr_init(&my_mutex_attr)	Initiates a mutex with attributes specified by **attr**. If **attr** is NULL, it applies the default mutex attributes.
pthread_mutexattr_destroy(&my_mutex_attr)	Destroys the specified mutex object. You can re-initialize with **pthread_mutex_init()**.
pthread_mutexattr_setrobust()	Specifies the robust attribute of a mutex.
pthread_mutexattr_getrobust()	Queries the robust attribute of a mutex .
PTHREAD_MUTEX_ROBUST	Defines a thread to terminate without unlocking the mutex. A future call to own this mutex succeeds automatically and returns the value **EOWNERDEAD** to indicate that the previous mutex owner no longer exists.
pthread_barrier_init()	Allocates the required resources to use and initialize the barrier with attribute object **attr**. If **attr** is NULL, it applies the default values.
pthread_cond_init()	Initializes a condition variable. The argument **cond** defines the object to initiate with the attributes in the condition variable attribute object **attr**. If **attr** is NULL, it applies the default values.

Function	Description
pthread_cond_wait()	Blocks a thread execution until it receives a signal from another thread. In addition, a call to this function also releases the associated lock on mutex before blocking. The argument **cond** defines the **pthread_cond_t** object for a thread to block on. The **mutex** argument specifies the mutex to unblock.
pthread_cond_signal()	Unblocks at least one of the threads that are blocked on a specified condition variable. The argument **cond** specifies using the **pthread_cond_t** object to unblock the thread.

CHAPTER 9. SOCKET OPTIONS IN RHEL FOR REAL TIME

A socket is a two way data transfer mechanism between two processes on same systems such as the UNIX domain and loopback devices or on different systems such as network sockets.

Transmission Control Protocol (TCP) is the most common transport protocol and is often used to achieve consistent low latency for a service that requires constant communication or to cork the sockets in a low priority restricted environment.

With new applications, hardware features, and kernel architecture optimizations, TCP has to introduce new approaches to handle the changes effectively. The new approaches can cause unstable program behaviors. Because the program behavior changes as the underlying operating system components change, they must be handled with care.

One example of such behavior in TCP is the delay in sending small buffers. This allows sending them as one network packet. Buffering small writes to TCP and sending them all at once generally works well, but it can also create latencies. For real-time applications, the **TCP_NODELAY** socket option disables the delay and sends small writes as soon as they are ready.

The relevant socket options for data transfer are **TCP_NODELAY** and **TCP_CORK**.

9.1. TCP_NODELAY SOCKET OPTION

The **TCP_NODELAY** socket option disables Nagle's algorithm. Configuring **TCP_NODELAY** with the **setsockopt** sockets API function sends multiple small buffer writes as individual packets as soon as they are ready.

Sending multiple logically related buffers as a single packet by building a contiguous packet before sending, achieves better latency and performance. Alternatively, if the memory buffers are logically related but not contiguous, you can create an I/O vector and pass it to the kernel using **writev** on a socket with **TCP_NODELAY** enabled.

The following example illustrates enabling **TCP_NODELAY** through the **setsockopt** sockets API.

```
# int one = 1;
# setsockopt(descriptor, SOL_TCP, TCP_NODELAY, &one, sizeof(one));
```

NOTE

To use **TCP_NODELAY** effectively, avoid small, logically related buffer writes. With **TCP_NODELAY**, small writes make TCP send multiple buffers as individual packets, which may result in poor overall performance.

Additional resources

- **sendfile(2)** man page

9.2. TCP_CORK SOCKET OPTION

The **TCP_CORK** option collects all data packets in a socket and prevents from transmitting them until the buffer fills to a specified limit. This enables applications to build a packet in the kernel space and send data when **TCP_CORK** is disabled. **TCP_CORK** is set on a socket file descriptor using the **setsocketopt()** function. When developing programs, if you must send bulk data from a file, consider using **TCP_CORK** with the **sendfile()** function.

When a logical packet is built in the kernel by various components, enable **TCP_CORK** by configuring it to a value of 1 using the **setsockopt** sockets API. This is known as "corking the socket". **TCP_CORK** can cause bugs if the cork is not removed at an appropriate time.

The following example illustrates enabling **TCP_CORK** through the **setsockopt** sockets API.

```
# int one = 1;
# setsockopt(descriptor, SOL_TCP, TCP_CORK, &one, sizeof(one));
```

In some environments, if the kernel is not able to identify when to remove the cork, you can manually remove it as follows:

```
# int zero = 0;
# setsockopt(descriptor, SOL_TCP, TCP_CORK, &zero, sizeof(zero));
```

Additional resources

- **sendfile(2)** man page

9.3. EXAMPLE PROGRAMS USING SOCKET OPTIONS

The **TCP_NODELAY** and **TCP_CORK** socket options significantly influence the behavior of a network connection. **TCP_NODELAY** disables the Nagle's algorithm on applications that benefit by sending data packets as soon as they are ready. With **TCP_CORK**, you can transfer multiple data packets together, with no delays between them.

The example programs in this section provide information on the performance impact these socket options can have on your applications.

Performance impact on a client

You can send small buffer writes on a client without using the **TCP_NODELAY** and **TCP_CORK** socket options. When run with no arguments, the client uses the default socket options.

- To initiate data transfer, define the server TCP port and the number of packets it must process. For example, 10,000 packets in this test.

  ```
  $ ./tcp_nodelay_server 5001 10000
  ```

  In all cases it sends 15 packets, each of two bytes, and waits for a response from the server.

Performance impact on a loopback interface

Following examples use a loopback interface to demonstrate three variations:

- To send buffer writes immediately, set the **no_delay** option on a socket configured with **TCP_NODELAY**.

  ```
  $ ./tcp_nodelay_client localhost 5001 10000 no_delay

   10000 packets of 30 bytes sent in 1649.771240 ms: 181.843399 bytes/ms using
  TCP_NODELAY
  ```

  TCP sends the buffers right away, disabling the algorithm that combines the small packets. This improves performance but can cause a flurry of small packets to be sent for each logical packet.

- To collect multiple data packets and send them with one system call, configure the **TCP_CORK** socket option.

```
$ ./tcp_nodelay_client localhost 5001 10000 cork

 10000 packets of 30 bytes sent in 850.796448 ms: 352.610779 bytes/ms using TCP_CORK
```

  Using the cork technique significantly reduces the time required to send data packets as it combines full logical packets in its buffers and sends fewer overall network packets. You must ensure to remove the **cork** at the appropriate time.

  When developing programs, if you must send bulk data from a file, consider using **TCP_CORK** with the **sendfile()** option.

- To measure performance without using socket options.

```
$ ./tcp_nodelay_client localhost 5001 10000

 10000 packets of 30 bytes sent in 400129.781250 ms: 0.749757 bytes/ms
```

  This is the baseline measure when TCP combines buffer writes and waits to check for more data than can optimally fit in the network packet.

Additional resources

- **sendfile(2)** man page

CHAPTER 10. RHEL FOR REAL TIME SCHEDULER

There are two different mechanisms to configure and monitor process configurations: the command line utilities, and the Tuna graphical tool. This section provides information on the command line tools. However, you can also perform all actions using the Tuna tool.

10.1. CHRT UTILITY FOR SETTING THE SCHEDULER

The **chrt** utility checks and adjusts scheduler policies and priorities. It can start new processes with the desired properties, or change the current properties of a running process.

The **chrt** utility takes the either **--pid** or the **-p** option to specify the process ID (PID).

The **chrt** utility takes the following policy options:

- **-f** or **--fifo**: sets the schedule to **SCHED_FIFO**.
- **-o** or **--other**: sets the schedule to **SCHED_OTHER**.
- **-r** or **--rr**: sets schedule to **SCHED_RR**.
- **-d** or **--deadline**: sets schedule to **SCHED_DEADLINE**.

The following example shows the attributes for a specified process.

```
# chrt -p 468
pid 468's current scheduling policy: SCHED_FIFO
pid 468's current scheduling priority: 85
```

10.2. PREEMPTIVE SCHEDULING

Preemption is the mechanism to temporarily interrupt an executing task, with the intention of resuming it at a later time. It occurs when a higher priority process interrupts to use the CPU. Preemption can have a particularly negative impact on performance, and constant preemption can lead to a state known as thrashing. This problem occurs when processes are constantly preempts and no process ever gets to run completely. Changing the priority of a task can help reduce involuntary preemption.

You can check for voluntary and involuntary preemption occurring on a single process by viewing the contents of the **/proc/PID/status** file, where PID is the process identifier.

The following example shows the preemption status of a process with PID 1000.

```
# grep voluntary /proc/1000/status
voluntary_ctxt_switches: 194529
nonvoluntary_ctxt_switches: 195338
```

10.3. LIBRARY FUNCTIONS FOR SCHEDULER PRIORITY

Real-time processes use a different set of library calls to control policy and priority. The functions require the inclusion of the **sched.h** header file. The symbols **SCHED_OTHER**, **SCHED_RR** and **SCHED_FIFO** must also be defined in the **sched.h** header file.

The following table lists the functions that set the policy and priority for the real-time scheduler.

Table 10.1. Library functions for real-time scheduler

functions	Description
sched_getscheduler()	Retrieves the scheduler policy for a specific process identifier (PID)
sched_setscheduler	Sets the scheduler policy and other parameters. This function requires three parameters: **sched_setscheduler(pid_t pid, int policy, const struct sched_param *sp);**
sched_getparam and sched_setparam	Sets the scheduling parameters of a particular process. This can then be verified using the **sched_getparam()** function.
sched_get_priority_min and sched_get_priority_max	Checks the valid priority range for a given scheduler policy.
sched_rr_get_interval	Reports the allocated **timeslice** for each process.

CHAPTER 11. SYSTEM CALLS IN RHEL FOR REAL TIME

System call is a function used by application programs to communicate with the kernel. It is a mechanism for programs to order resources from the kernel.

11.1. SCHED_YIELD FUNCTION

The **sched_yield** function is designed for a processor to select a process other than the running one. This type of request is prone to failure when issued from within a poorly-written application.

When the **sched_yield()** function is used within processes with real-time priorities, it can display unexpected behavior. The process that calls **sched_yield** moves to the tail of the queue of processes running at same priority. When there are no other processes running at the same priority, the process that called **sched_yield()** continues to run. If the priority of that process is high, it can potentially create a busy loop, rendering the machine unusable.

In general, do not use **sched_yield** on real-time processes.

11.2. GETRUSAGE FUNCTION

The **getrusage** function retrieves important information from a specified process or its threads. It reports on information such as:

- The number of voluntary and involuntary context switches.
- Major and minor page faults.
- Amount of memory in use.

getrusage enables you to query an application to provide information relevant to both performance tuning and debugging activities. **getrusage()** retrieves information that would otherwise need to be cataloged from several different files in the /**proc**/ directory and would be hard to synchronize with specific actions or events on the application.

NOTE

Not all the fields contained in the structure filled with **getrusage()** results are set by the kernel. Some of them are kept for compatibility reasons only.

Additional resources

- **getrusage(2)** man page

CHAPTER 12. USING MLOCK SYSTEM CALLS ON RHEL FOR REAL TIME

The memory lock (**mlock**) function enables the real-time calling processes to lock or unlock a specified range of the address space. This prevents Linux from paging the locked memory when swapping memory space. Once you allocate the physical page to the page table entry, references to that page will always be fast. The **mlock()** system calls include two functions: **mlock()** and **mlockall()**. Similarly, **munlock()** system call includes the **munlock()** and **munlockall()** functions.

12.1. MLOCK AND MUNLOCK SYSTEM CALLS

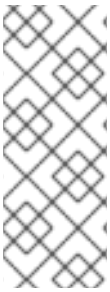
The **mlock()** and **mlockall()** system calls lock a specified memory range and does not allow that memory to be paged. This means that once the physical page is allocated to the page table entry, references to that page will always be fast.

There are two groups of **mlock()** system calls: **mlock()** and **munlock()** The **mlock()** and **munlock()** calls lock and unlock a specific range of addresses. **mlock()** system calls lock pages in the address range starting at **addr** and continuing for **len** bytes. When the call returns successfully, all pages that contain a part of the specified address range stay in the memory until unlocked later.

With **mlockall()**, you can lock all mapped pages into the specified address range. These can be pages of code, shared memory, memory mapped files, and so on. Memory locks do not stack. This means that any page locked by several calls will unlock the specified address range or the entire region with a single **munlock()** call. With **munlockall()** calls, you can unlock the entire program space.

The status of the pages contained in a specific range depends on the value in the **flags** argument. The **flags** argument can be either 0 or **MLOCK_ONFAULT**.

Memory locks are not inherited by a child process via fork and are removed automatically when a process terminates.



NOTE

Use **mlock()** with caution. Excessive use can cause out-of-memory (OOM) errors. When an application is large or if it has a large data domain, the **mlock()** calls can cause thrashing when the system is not able to allocate memory for other tasks.

When using **mlockall()** calls for real-time processes, ensure that you reserve sufficient stack pages.

12.2. USING MLOCK SYSTEM CALLS TO LOCK PAGES

mlock() calls use the **addr** parameter to specify the start of an address range and **len** to define the length of the address space in bytes.

The function **alloc_workbuf** dynamically allocates a memory buffer and locks it. The memory allocation is performed by the **posix_memalig** function in order to align the memory area to a page. The function **free_workbuf** unlocks the memory area.

Prerequisites:

- Root privileges.

- In case of insufficient privileges, you must have the **CAP_IPC_LOCK** capability to use **mlockall()** or **mlock()** on large buffers.

Procedure

- To lock pages using **mlock()**, run the following command:

```
#include <stdlib.h>
#include <unistd.h>
#include <sys/mman.h>

void *alloc_workbuf(size_t size)
{
  void ptr;
  int retval;

  // alloc memory aligned to a page, to prevent two mlock() in the same page.
  retval = posix_memalign(&ptr, (size_t) sysconf(_SC_PAGESIZE), size);

  // return NULL on failure
  if (retval)
    return NULL;

  // lock this buffer into RAM
  if (mlock(ptr, size)) {
    free(ptr);
    return NULL;
  }
    return ptr;
}

void free_workbuf(void *ptr, size_t size) {
  // unlock the address range
  munlock(ptr, size);

  // free the memory
  free(ptr);
}
```

Verification

Upon successful completion, **mlock()** and **munlock()** return 0. In case of an error, they return -1 and set a **errno** to indicate the error.

12.3. USING MLOCKALL SYSTEM CALLS TO LOCK ALL MAPPED PAGES

To lock and unlock memory with **mlockall()** and **munlockall()**, set the **flags** argument as either 0 or one of the constant: **MCL_CURRENT** or **MCL_FUTURE**. With **MCL_FUTURE**, a future system call, such as **mmap2()**, **sbrk2()**, or **malloc3()**, might fail, as it causes the number of locked bytes to exceed the permitted maximum.

Prerequisites

- Root privileges

Procedure

- To lock all mapped pages using **mlockall()**:

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

 int mlockall (init flags)
```

- To unlock all mapped pages using **munlockall()**:

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

 int munlockall (void)
```

Additional resources

- **capabilities(7)** man page
- **mlock(2)** man page
- **mlock(3)** man page
- **move_pages(2)** man page
- **posix_memalign(3)** man page
- **posix_memalign(3p)** man page

12.4. USING MMAP SYSTEM CALLS TO MAP FILES OR DEVICES INTO MEMORY

For large memory allocations, the memory allocation (**malloc**) method uses the **mmap()** system call to find addressable memory space. You can allocate and lock memory areas by setting **MAP_LOCKED** in the **flags** parameter.

As **mmap()** allocates memory on a page basis, there are no two locks on the same page, which prevents the double-lock or single-unlock problems.

Prerequisites

- Root privileges

Procedure

- To map a specific process address space, use the following command:

```
#include <sys/mman.h>
#include <stdlib.h>

void *alloc_workbuf(size_t size)
{
 void *ptr;

 ptr = mmap(NULL, size, PROT_READ | PROT_WRITE,
```

```
            MAP_PRIVATE | MAP_ANONYMOUS | MAP_LOCKED, -1, 0);

 if (ptr == MAP_FAILED)
  return NULL;

 return ptr;
}

void
free_workbuf(void *ptr, size_t size)
{
 munmap(ptr, size);
}
```

Verification

Upon success, **mmap()** returns a pointer to the mapped area. On error, it returns the value **MAP_FAILED** and sets a **errno** to indicate the error.

Upon success, **munmap()** returns 0. On error, it returns -1 and sets a **errno** to indicate the error.

Additional resources

- **mmap(2)** man page
- **mlockall(2)** man page

12.5. PARAMETERS FOR MLOCK SYSTEM CALLS

The following table lists the **mlock()** parameters.

Table 12.1. mlock parameters

Parameter	Description
addr	Specifies the process address space to lock or unlock. When NULL, the kernel chooses the page-aligned arrangement of data in the memory. If **addr** is not NULL, the kernel chooses a nearby page boundary, which is always above or equal to the value specified in **/proc/sys/vm/mmap_min_addr** file.
len	Specifies the length of the mapping, which must be greater than 0.
fd	Specifies the file descriptor.
prot	**mmap** and **munmap** calls define the desired memory protection with this parameter. **prot** takes one or a combination of **PROT_EXEC**, **PROT_READ**, **PROT_WRITE** or **PROT_NONE** values.

Parameter	Description
flags	Controls the mapping visibility to other processes that map the same file. It takes one of the values: **MAP_ANONYMOUS**, **MAP_LOCKED**, **MAP_PRIVATE** or **MAP_SHARED** values.
MCL_CURRENT	Locks all pages that are currently mapped into a process.
MCL_FUTURE	Sets the mode to lock subsequent memory allocations. These could be new pages required by a growing heap and stack, new memory-mapped files, or shared memory regions.

CHAPTER 13. SETTING CPU AFFINITY ON RHEL FOR REAL TIME

Each thread and interrupt source in the system has a processor affinity property. The operating system scheduler uses this information to determine the threads and interrupts to run on a CPU. Setting processor affinity, along with effective policy and priority settings, can help to achieve the maximum possible performance. Applications will always need to compete for resources, especially CPU time, with other processes. Depending on the application, related threads are often run on the same core. Alternatively, one application thread can be allocated to one core.

Systems that perform multitasking are naturally more prone to indeterminism. Even high priority applications may be delayed from executing while a lower priority application is in a critical section of code. Once the low priority application has exited the critical section, the kernel may safely preempt the low priority application and schedule the high priority application on the processor. Additionally, migrating processes from one CPU to another can be costly due to cache invalidation. Red Hat Enterprise Linux for Real Time includes tools that address some of these issues and allow latencies to be better controlled.

Affinity is represented as a bitmask, where each bit in the mask represents a CPU core. If the bit is set to 1, then the thread or interrupt may run on that core; if 0 then the thread or interrupt is excluded from running on the core. The default value for an affinity bitmask is all ones, meaning the thread or interrupt may run on any core in the system.

By default, processes can run on any CPU. However, by changing the affinity of the process, you can define a process to run on a predetermined set of CPUs. Child processes inherit the CPU affinities of their parents.

Setting the following typical affinity setups can achieve maximum possible performance:

- Using a single CPU core for all system processes and setting the application to run on the remainder of the cores.
- Configuring a thread application and a specific kernel thread (network softirq or a driver thread) on the same CPU.
- Pairing the producer-consumer threads on each CPU. Producers and consumers are two classes of threads, where producers insert data into the buffer and consumers remove it from the buffer.

The usual good practice for tuning affinities on a real-time system is to determine the number of cores required to run the application and then isolate those cores. You achieve this with the Tuna tool or with the shell scripts to modify the bitmask value, such as the **taskset** command. The **taskset** command changes the affinity of a process and modifying the /**proc**/ file system entry changes the affinity of an interrupt.

13.1. TUNING PROCESSOR AFFINITY USING THE TASKSET COMMAND

The **taskset** command helps to set or retrieve the CPU affinity of a running process. The **taskset** command takes **-p** and **-c** options. The **-p** or **--pid** option work an existing process and does not start a new task. The **-c** or **--cpu-list** specify a numerical list of processors instead of a bitmask. The list may contain multiple items, separated by comma, and a range of processors. For example, 0,5,7,9-11.

Prerequisites

- Root privileges

Procedure

- To check the process affinity for a specific process, run the following command:

```
# taskset -p -c 1000
  pid 1000's current affinity list: 0,1
```

The command prints the affinity of the process with PID 1000. The process is configured to use either CPU 0 or CPU 1.

- (Optional) To configure a specific CPU to bind a process, run the following command:

```
# taskset -p -c 1 1000
pid 1000's current affinity list: 0,1
pid 1000's new affinity list: 1
```

- (Optional) To define more than one CPU affinity, run the following command:

```
# taskset -p -c 0,1 1000
pid 1000's current affinity list: 1
pid 1000's new affinity list: 0,1
```

- (Optional) To configure a priority level and a policy on a specific CPU, run the following command:

```
# taskset -c 5 chrt -f 78 /bin/my-app
```

For further granularity, you can also specify the priority and policy. In the example, the command runs the **/bin/my-app** application on CPU 5 with **SCHED_FIFO** policy and a priority value of 78.

13.2. SETTING PROCESSOR AFFINITY USING THE SCHED_SETAFFINITY() SYSTEM CALL

In addition to the **taskset** command, you can also set processor affinity using the **sched_setaffinity()** system call.

Prerequisite

- Root privileges

Procedure

- To set the processor affinity with **sched_setaffinity()**, run the following command:

```
#define _GNU_SOURCE
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <errno.h>
#include <sched.h>

int main(int argc, char **argv)
{
```

```
  int i, online=0;
  ulong ncores = sysconf(_SC_NPROCESSORS_CONF);
  cpu_set_t *setp = CPU_ALLOC(ncores);
  ulong setsz = CPU_ALLOC_SIZE(ncores);

  CPU_ZERO_S(setsz, setp);

  if (sched_getaffinity(0, setsz, setp) == -1) {
    perror("sched_getaffinity(2) failed");
    exit(errno);
  }

  for (i=0; i < CPU_COUNT_S(setsz, setp); i) {
    if (CPU_ISSET_S(i, setsz, setp))
      online;
  }

  printf("%d cores configured, %d cpus allowed in affinity mask\n", ncores, online);
  CPU_FREE(setp);
}
```

13.3. ISOLATING A SINGLE CPU TO RUN HIGH UTILIZATION TASKS

Using the **cpusets** mechanism, you can assign a set of CPUs and memory nodes for **SCHED_DEADLINE** tasks.

In a task set which includes high and low CPU utilizing tasks, isolating a CPU to run the high utilization task and scheduling small utilization tasks on different sets of CPU, enables all tasks to meet the assigned **runtime**.

Prerequisites

- Root privileges

Procedure

1. Create two directories named as **cpuset**:

```
# cd /sys/fs/cgroup/cpuset/
# mkdir cluster
# mkdir partition
```

2. Disable the load balance of the root **cpuset** to create two new root domains in the **cpuset** directory:

```
# echo 0 > cpuset.sched_load_balance
```

3. In the cluster **cpuset**, schedule the low utilization tasks to run on CPU 1 to 7, verify memory size, and name the CPU as exclusive:

```
# cd cluster/
# echo 1-7 > cpuset.cpus
# echo 0 > cpuset.mems
# echo 1 > cpuset.cpu_exclusive
```

4. Move all low utilization tasks to the cpuset directory:

```
# ps -eLo lwp | while read thread; do echo $thread > tasks ; done
```

5. Create a partition named as **cpuset** and assign the high utilization task:

```
# cd ../partition/
# echo 1 > cpuset.cpu_exclusive
# echo 0 > cpuset.mems
# echo 0 > cpuset.cpus
```

6. Set the shell to the cpuset and start the deadline workload:

```
# echo $$ > tasks
# /root/d &
```

With this setup, the task isolated in the partitioned **cpuset** directory does not interfere with the task in the cluster **cpuset** directory. This enables all real-time tasks to meet the scheduler deadline.