Fedora 15

Power Management Guide

Managing power consumption on Fedora

Don Domingo

Rüdiger Landmann

Jack Reed

Fedora 15 Power Management Guide
Managing power consumption on Fedora
Edition 1.0

Author	Don Domingo	
Author	Rüdiger Landmann	
Author	Jack Reed	jreed@redhat.com

This document explains how to manage power consumption on Fedora 15 systems effectively. The following sections discuss different techniques that lower power consumption (for both server and laptop), and how each technique affects the overall performance of your system. Please note: This document is still under development, is subject to heavy change, and is provided here as a preview. The content and instructions contained within should not be considered complete, and should be used with caution.

Preface

1. Document Conventions

This manual uses several conventions to highlight certain words and phrases and draw attention to specific pieces of information.

In PDF and paper editions, this manual uses typefaces drawn from the *Liberation Fonts*[1] set. The Liberation Fonts set is also used in HTML editions if the set is installed on your system. If not, alternative but equivalent typefaces are displayed. Note: Red Hat Enterprise Linux 5 and later includes the Liberation Fonts set by default.

1.1. Typographic Conventions

Four typographic conventions are used to call attention to specific words and phrases. These conventions, and the circumstances they apply to, are as follows.

`Mono-spaced Bold`

Used to highlight system input, including shell commands, file names and paths. Also used to highlight keycaps and key combinations. For example:

> To see the contents of the file **`my_next_bestselling_novel`** in your current working directory, enter the **`cat my_next_bestselling_novel`** command at the shell prompt and press **`Enter`** to execute the command.

The above includes a file name, a shell command and a keycap, all presented in mono-spaced bold and all distinguishable thanks to context.

Key combinations can be distinguished from keycaps by the hyphen connecting each part of a key combination. For example:

> Press **`Enter`** to execute the command.
>
> Press **`Ctrl`**+**`Alt`**+**`F2`** to switch to the first virtual terminal. Press **`Ctrl`**+**`Alt`**+**`F1`** to return to your X-Windows session.

The first paragraph highlights the particular keycap to press. The second highlights two key combinations (each a set of three keycaps with each set pressed simultaneously).

If source code is discussed, class names, methods, functions, variable names and returned values mentioned within a paragraph will be presented as above, in **`mono-spaced bold`**. For example:

> File-related classes include **`filesystem`** for file systems, **`file`** for files, and **`dir`** for directories. Each class has its own associated set of permissions.

Proportional Bold

This denotes words or phrases encountered on a system, including application names; dialog box text; labeled buttons; check-box and radio button labels; menu titles and sub-menu titles. For example:

> Choose **System** → **Preferences** → **Mouse** from the main menu bar to launch **Mouse Preferences**. In the **Buttons** tab, click the **Left-handed mouse** check box and click

[1] https://fedorahosted.org/liberation-fonts/

Close to switch the primary mouse button from the left to the right (making the mouse suitable for use in the left hand).

To insert a special character into a **gedit** file, choose **Applications** → **Accessories** → **Character Map** from the main menu bar. Next, choose **Search** → **Find…** from the **Character Map** menu bar, type the name of the character in the **Search** field and click **Next**. The character you sought will be highlighted in the **Character Table**. Double-click this highlighted character to place it in the **Text to copy** field and then click the **Copy** button. Now switch back to your document and choose **Edit** → **Paste** from the **gedit** menu bar.

The above text includes application names; system-wide menu names and items; application-specific menu names; and buttons and text found within a GUI interface, all presented in proportional bold and all distinguishable by context.

Mono-spaced Bold Italic or ***Proportional Bold Italic***

Whether mono-spaced bold or proportional bold, the addition of italics indicates replaceable or variable text. Italics denotes text you do not input literally or displayed text that changes depending on circumstance. For example:

To connect to a remote machine using ssh, type **ssh *username*@*domain.name*** at a shell prompt. If the remote machine is **example.com** and your username on that machine is john, type **ssh john@example.com**.

The **mount -o remount *file-system*** command remounts the named file system. For example, to remount the **/home** file system, the command is **mount -o remount /home**.

To see the version of a currently installed package, use the **rpm -q *package*** command. It will return a result as follows: ***package-version-release***.

Note the words in bold italics above — username, domain.name, file-system, package, version and release. Each word is a placeholder, either for text you enter when issuing a command or for text displayed by the system.

Aside from standard usage for presenting the title of a work, italics denotes the first use of a new and important term. For example:

Publican is a *DocBook* publishing system.

1.2. Pull-quote Conventions

Terminal output and source code listings are set off visually from the surrounding text.

Output sent to a terminal is set in **mono-spaced roman** and presented thus:

```
books        Desktop   documentation  drafts  mss    photos   stuff  svn
books_tests  Desktop1  downloads      images  notes  scripts  svgs
```

Source-code listings are also set in **mono-spaced roman** but add syntax highlighting as follows:

```
package org.jboss.book.jca.ex1;

import javax.naming.InitialContext;
```

```
public class ExClient
{
   public static void main(String args[])
       throws Exception
   {
      InitialContext iniCtx = new InitialContext();
      Object         ref    = iniCtx.lookup("EchoBean");
      EchoHome       home   = (EchoHome) ref;
      Echo           echo   = home.create();

      System.out.println("Created Echo");

      System.out.println("Echo.echo('Hello') = " + echo.echo("Hello"));
   }
}
```

1.3. Notes and Warnings

Finally, we use three visual styles to draw attention to information that might otherwise be overlooked.

Note

Notes are tips, shortcuts or alternative approaches to the task at hand. Ignoring a note should have no negative consequences, but you might miss out on a trick that makes your life easier.

Important

Important boxes detail things that are easily missed: configuration changes that only apply to the current session, or services that need restarting before an update will apply. Ignoring a box labeled 'Important' will not cause data loss but may cause irritation and frustration.

Warning

Warnings should not be ignored. Ignoring warnings will most likely cause data loss.

2. We Need Feedback!

If you find a typographical error in this manual, or if you have thought of a way to make this manual better, we would love to hear from you! Please submit a report in Bugzilla: *http://bugzilla.redhat.com/bugzilla/* against the product **Fedora Documentation.**

When submitting a bug report, be sure to mention the manual's identifier: *power-management-guide*

If you have a suggestion for improving the documentation, try to be as specific as possible when describing it. If you have found an error, please include the section number and some of the surrounding text so we can find it easily.

Overview

Limiting the power used by computer systems is one of the most important aspects of *green IT* (environmentally friendly computing), a set of considerations that also encompasses the use of recyclable materials, the environmental impact of hardware production, and environmental awareness in the design and deployment of systems. In this document, we provide guidance and information regarding power management of your systems running Fedora 15.

1.1. Importance of Power Management

At the core of power management is an understanding of how to effectively optimize energy consumption of each system component. This entails studying the different tasks that your system performs, and configuring each component to ensure that its performance is just right for the job.

The main motivator for power management is:

- reducing overall power consumption to save cost

The proper use of power management results in:

- heat reduction for servers and computing centers
- reduced secondary costs, including cooling, space, cables, generators, and *uninterruptible power supplies* (UPS)
- extended battery life for laptops
- lower carbon dioxide output
- meeting government regulations or legal requirements regarding Green IT, for example Energy Star
- meeting company guidelines for new systems

As a rule, lowering the power consumption of a specific component (or of the system as a whole) will lead to lower heat and naturally, performance. As such, you should thoroughly study and test the decrease in performance afforded by any configurations you make, especially for mission-critical systems.

By studying the different tasks that your system performs, and configuring each component to ensure that its performance is just sufficient for the job, you can save energy, generate less heat, and optimize battery life for laptops. Many of the principles for analysis and tuning of a system in regard to power consumption are similar to those for performance tuning. To some degree, power management and performance tuning are opposite approaches to system configuration, because systems are usually optimized either towards performance or power. This manual describes the tools that the Fedora Project provides and the techniques we have developed to help you in this process.

Fedora 15 already comes with a lot of new power management features that are enabled by default. They were all selectively chosen to not impact the performance of a typical server or desktop use case. However, for very specific use cases where maximum throughput, lowest latency, or highest CPU performance is absolutely required, a review of those defaults might be necessary.

To decide whether you should optimize your machines using the techniques described in this document, ask yourself a few questions:

Q: Must I optimize?

A: The importance of power optimization depends on whether your company has guidelines that need to be followed or if there are any regulations that you have to fulfill.

Q: How much do I need to optimize?

A: Several of the techniques we present do not require you to go through the whole process of auditing and analyzing your machine in detail but instead offer a set of general optimizations that typically improve power usage. Those will of course typically not be as good as a manually audited and optimized system, but provide a good compromise.

Q: Will optimization reduce system performance to an unacceptable level?

A: Most of the techniques described in this document impact the performance of your system noticeably. If you choose to implement power management beyond the defaults already in place in Fedora 15, you should monitor the performance of the system after power optimization and decide if the performance loss is acceptable.

Q: Will the time and resources spent to optimize the system outweigh the gains achieved?

A: Optimizing a single system manually following the whole process is typically not worth it as the time and cost spent doing so is far higher than the typical benefit you would get over the lifetime of a single machine. On the other hand if you for example roll out 10000 desktop systems to your offices all using the same configuration and setup then creating one optimized setup and applying that to all 10000 machines is most likely a good idea.

The following sections will explain how optimal hardware performance benefits your system in terms of energy consumption.

1.2. Power Management Basics

Effective power management is built on the following principles:

An idle CPU should only wake up when needed

The kernel used in releases of Fedora prior to Fedora 8 used a periodic timer for each CPU. This timer prevents the CPU from truly going idle, as it requires the CPU to process each timer event (which would happen every few milliseconds, depending on the setting), regardless of whether any process was running or not. A large part of effective power management involves reducing the frequency at which CPU wakeups are made.

Because of this, the Linux kernel in Fedora 15 eliminates the periodic timer: as a result, the idle state of a CPU is now *tickless*. This prevents the CPU from consuming unnecessary power when it is idle. However, benefits from this feature can be offset if your system has applications that create unnecessary timer events. Polling events (such as checks for volume changes, mouse movement, and the like) are examples of such events.

Fedora 15 includes tools with which you can identify and audit applications on the basis of their CPU usage. Refer to *Chapter 2, Power management auditing and analysis* for details.

Unused hardware and devices should be disabled completely

This is especially true for devices that have moving parts (for example, hard disks). In addition to this, some applications may leave an unused but enabled device "open"; when this occurs, the kernel assumes that the device is in use, which can prevent the device from going into a power saving state.

Low activity should translate to low wattage

In many cases, however, this depends on modern hardware and correct BIOS configuration. Older system components often do not have support for some of the new features that we now can support in Fedora 15. Make sure that you are using the latest official firmware for your systems and that in the power management or device configuration sections of the BIOS the power management features are enabled. Some features to look for include:

- SpeedStep
- PowerNow!
- Cool'n'Quiet
- ACPI (C state)
- Smart

If your hardware has support for these features and they are enabled in the BIOS, Fedora 15 will use them by default.

Different forms of CPU states and their effects

Modern CPUs together with *Advanced Configuration and Power Interface* (ACPI) provide different power states. The three different states are:

- Sleep (C-states)
- Frequency (P-states)
- Heat output (T-states or "thermal states")

A CPU running on the lowest sleep state possible (indicated by the highest C-state number) consumes the least amount of watts, but it also takes considerably more time to wake it up from that state when needed. In very rare cases this can lead to the CPU having to wake up immediately every time it just went to sleep. This situation results in an effectively permanently busy CPU and loses some of the potential power saving if another state had been used.

A turned off machine uses the least amount of power

As obvious as this might sound, one of the best ways to actually save power is to turn off systems. For example, your company can develop a corporate culture focused on "green IT" awareness with a guideline to turn of machines during lunch break or when going home. You also might consolidate several physical servers into one bigger server and virtualize them using the virtualization technology we ship with Fedora 15.

Power management auditing and analysis

2.1. Audit and analysis overview

The detailed manual audit, analysis, and tuning of a single system is usually the exception because the time and cost spent to do so typically outweighs the benefits gained from these last pieces of system tuning. However, performing these tasks once for a large number of nearly identical systems where you can reuse the same settings for all systems can be very useful. For example, consider the deployment of thousands of desktop systems, or a HPC cluster where the machines are nearly identical. Another reason to do auditing and analysis is to provide a basis for comparison against which you can identify regressions or changes in system behavior in the future. The results of this analysis can be very helpful in cases where hardware, BIOS, or software updates happen regularly and you want to avoid any surprises with regard to power consumption. Generally, a thorough audit and analysis gives you a much better idea of what is really happening on a particular system.

Auditing and analyzing a system with regard to power consumption is relatively hard, even with the most modern systems available. Most systems do not provide the necessary means to measure power use via software. Exceptions exist though: the ILO management console of Hewlett Packard server systems has a power management module that you can access through the web. IBM provides a similar solution in their BladeCenter power management module. On some Dell systems, the IT Assistant offers power monitoring capabilities as well. Other vendors are likely to offer similar capabilities for their server platforms, but as can be seen there is no single solution available that is supported by all vendors. If your system has no inbuilt mechanism to measure power consumption, a few other choices exist. You could install a special power supply for your system that offers power consumption information through USB. The Gigabyte Odin GT 550 W PC power supply is one such example, and software to read out those values under Linux is available externally from *http://mgmt.sth.sze.hu/~andras/dev/gopsu/*. As a last resort, some external watt meters like the Watts up? PRO have a USB connector.

Direct measurements of power consumption is often only necessary to maximize savings as far as possible. Fortunately, other means are available to measure if changes are in effect or how the system is behaving. This chapter describes the necessary tools.

2.2. PowerTOP

The tickless kernel in Fedora allows the CPU to enter the idle state more frequently, reducing power consumption and improving power management. The new **PowerTOP** tool identifies specific components of kernel and user-space applications that frequently wake up the CPU. **PowerTOP** was used in development to perform the audits described in *Section 3.11, “Optimizations in User Space”* that led to many applications being tuned in this release, reducing unnecessary CPU wake up by a factor of ten.

Fedora 15 comes with version 2.x of **PowerTOP**. This version is a complete rewrite of the 1.x code base. It features a clearer tab-based user interface and extensively uses the kernel "perf" infrastructure to give more accurate data. The power behavior of system devices is tracked and prominently displayed, so problems can be pinpointed quickly. More experimentally, the 2.x codebase includes a power estimation engine that can indicate how much power individual devices and processes are consuming. Refer to *Figure 2.1, “PowerTOP in Operation”*.

Install **PowerTOP** with the command:

```
yum install powertop
```

PowerTOP can provide an estimate of the total power usage of the system and show individual power usage for each process, device, kernel work, timer, and interrupt handler. Laptops should run on battery power during this task. To calibrate the power estimation engine, run the command:

```
powertop --calibrate
```

Calibration takes time and requires root privileges. The process performs various tests, and will cycle through brightness levels and switch devices on and off. Do not touch the machine during the calibration. When the calibration process finishes, **PowerTOP** starts as normal. Let it run for approximately an hour to collect data. When enough data is collected, power estimation figures will begin appearing in the first column.

Next, run **PowerTOP** with the command:

```
powertop
```

You will need to run **PowerTOP** with root privileges. Your laptop should still be running on battery power so that all available data will be presented.

While it runs, **PowerTOP** gathers statistics from the system. In the **Overview** tab, you can view a list of the components that are either sending wake-ups to the CPU most frequently or are consuming the most power (refer to *Figure 2.1, "PowerTOP in Operation"*). The adjacent columns display power estimation, how the resource is being used, wakeups per second, the classification of the component (such as process, device, or timer), and a description of the component. Wakeups per second indicates how efficiently the services or the devices and drivers of the kernel are performing. Less wakeups means less power is consumed. Components are ordered by how much further their power usage can be optimized.

Tuning driver components typically requires kernel changes, which is beyond the scope of this document. However, userland processes that send wakeups are more easily managed. First, identify if this service or application needs to run at all on this system. If not, simply deactivate it. To turn off an old SYSV service permanently, run:

```
chkconfig servicename off
```

To turn off a new systemd service permanently, run:

```
systemctl servicename.service disable
```

If you need more details about the what the component actually does, run:

```
ps -awux | grep componentname
strace -p processid
```

If the trace looks like it is repeating itself, then you probably have found a busy loop. Fixing such bugs typically requires a code change in that component, which again goes beyond the scope of this document.

As seen in *Figure 2.1, "PowerTOP in Operation"*, total power consumption and the remaining battery life are displayed, if applicable. Below these is a short summary, featuring total wakeups per second, GPU operations per second, and virtual filesystem operations per second.

Use the **left** and **right** keys to cycle through tabs. In the **Idle stats** tab, use of C-states is shown for all processors and cores. In the **Frequency stats** tab, use of P-states including the Turbo mode (if applicable) is shown for all processors and cores. The longer the CPU stays in the higher C- or P-states, the better (**C4** being higher than **C3**). This is a good indication of how well CPU usage has been optimized. Residency should ideally be 90% or more in the highest C- or P-state while the system is idle.

The **Device Stats** tab provides similar information to **Overview** but only for devices.

The **Tunables** tab contains suggestions for optimizing the system for lower power consumption. The **up** and **down** keys can be used to move through suggestions. The **enter** key toggles the suggestion on and off. These tunings are not persistent across reboots. This problem has been referred to developers.

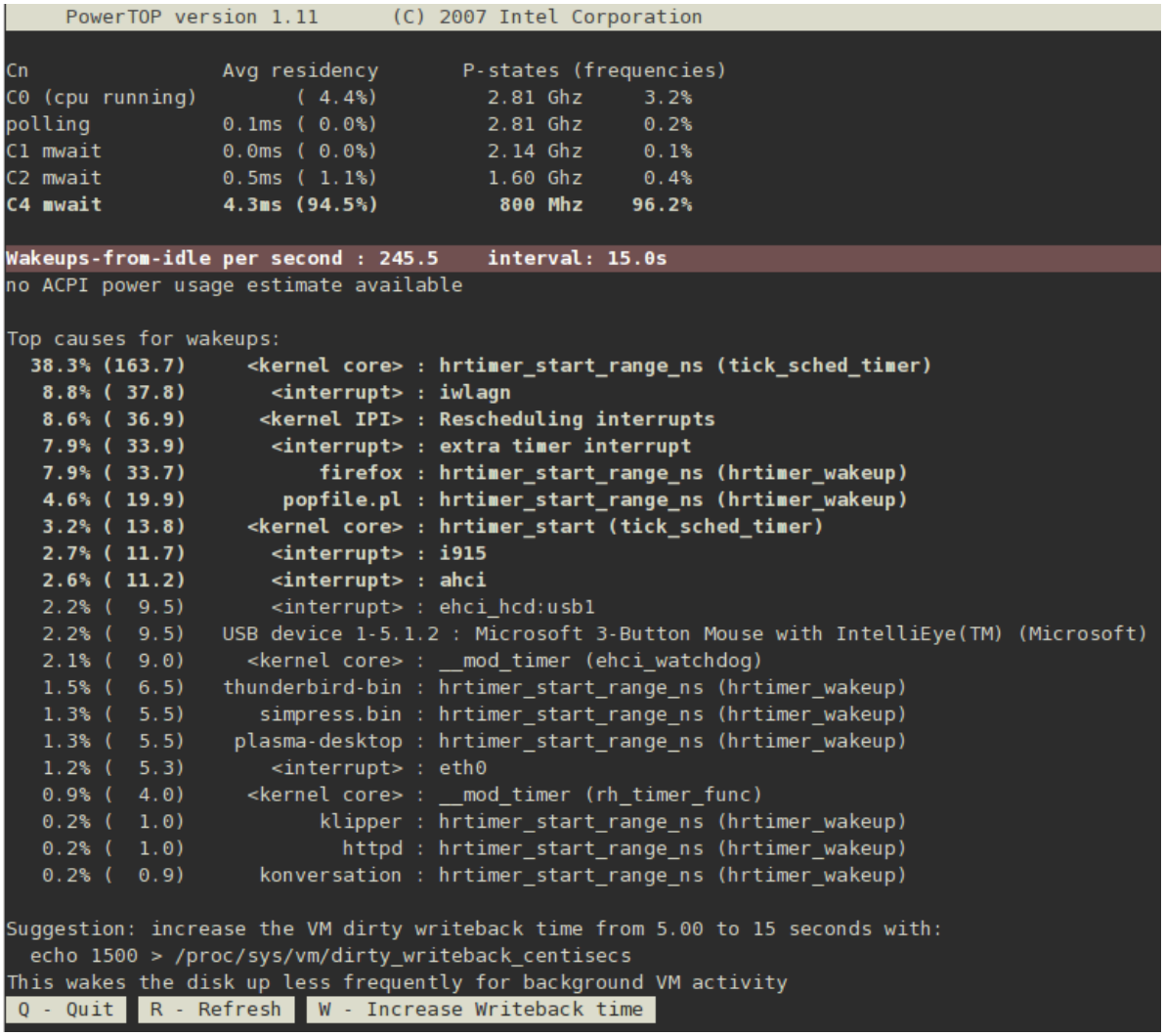

Figure 2.1. PowerTOP in Operation

The *Less Watts* website publishes a list of applications that **PowerTOP** has identified as keeping CPUs active. Refer to *http://www.lesswatts.org/projects/powertop/known.php*.

2.3. Diskdevstat and netdevstat

Diskdevstat and **netdevstat** are **SystemTap** tools that collect detailed information about the disk activity and network activity of all applications running on a system. These tools were inspired by

PowerTOP, which shows the number of CPU wakeups by every application per second (refer to *Section 2.2, "PowerTOP"*). The statistics that these tools collect allow you to identify applications that waste power with many small I/O operations rather than fewer, larger operations. Other monitoring tools that measure only transfer speeds do not help to identify this type of usage.

Install these tools with **SystemTap** with the command:

```
yum install systemtap tuned-utils kernel-debuginfo
```

Run the tools with the command:

```
diskdevstat
```

or the command:

```
netdevstat
```

Both commands can take up to three parameters, as follows:

diskdevstat ***update_interval total_duration display_histogram***

netdevstat ***update_interval total_duration display_histogram***

update_interval
: The time in seconds between updates of the display. Default: **5**

total_duration
: The time in seconds for the whole run. Default: **86400** (1 day)

display_histogram
: Flag whether to histogram for all the collected data at the end of the run.

The output resembles that of **PowerTOP**. Here is sample output from a longer **diskdevstat** run on a Fedora 10 system running KDE 4.2:

```
  PID   UID DEV     WRITE_CNT WRITE_MIN WRITE_MAX WRITE_AVG    READ_CNT  READ_MIN  READ_MAX
 READ_AVG COMMAND
 2789  2903 sda1          854     0.000   120.000    39.836           0     0.000     0.000
   0.000 plasma
15494     0 sda1            0     0.000     0.000     0.000         758     0.000     0.012
   0.000 0logwatch
15520     0 sda1            0     0.000     0.000     0.000         140     0.000     0.009
   0.000 perl
15549     0 sda1            0     0.000     0.000     0.000         140     0.000     0.009
   0.000 perl
15585     0 sda1            0     0.000     0.000     0.000         108     0.001     0.002
   0.000 perl
 2573     0 sda1           63     0.033  3600.015   515.226           0     0.000     0.000
   0.000 auditd
15429     0 sda1            0     0.000     0.000     0.000          62     0.009     0.009
   0.000 crond
15379     0 sda1            0     0.000     0.000     0.000          62     0.008     0.008
   0.000 crond
15473     0 sda1            0     0.000     0.000     0.000          62     0.008     0.008
   0.000 crond
15415     0 sda1            0     0.000     0.000     0.000          62     0.008     0.008
   0.000 crond
15433     0 sda1            0     0.000     0.000     0.000          62     0.008     0.008
   0.000 crond
15425     0 sda1            0     0.000     0.000     0.000          62     0.007     0.007
   0.000 crond
```

```
15375     0 sda1            0     0.000     0.000     0.000          62     0.008     0.008
   0.000 crond
15477     0 sda1            0     0.000     0.000     0.000          62     0.007     0.007
   0.000 crond
15469     0 sda1            0     0.000     0.000     0.000          62     0.007     0.007
   0.000 crond
15419     0 sda1            0     0.000     0.000     0.000          62     0.008     0.008
   0.000 crond
15481     0 sda1            0     0.000     0.000     0.000          61     0.000     0.001
   0.000 crond
15355     0 sda1            0     0.000     0.000     0.000          37     0.000     0.014
   0.001 laptop_mode
 2153     0 sda1           26     0.003  3600.029  1290.730           0     0.000     0.000
   0.000 rsyslogd
15575     0 sda1            0     0.000     0.000     0.000          16     0.000     0.000
   0.000 cat
15581     0 sda1            0     0.000     0.000     0.000          12     0.001     0.002
   0.000 perl
15582     0 sda1            0     0.000     0.000     0.000          12     0.001     0.002
   0.000 perl
15579     0 sda1            0     0.000     0.000     0.000          12     0.000     0.001
   0.000 perl
15580     0 sda1            0     0.000     0.000     0.000          12     0.001     0.001
   0.000 perl
15354     0 sda1            0     0.000     0.000     0.000          12     0.000     0.170
   0.014 sh
15584     0 sda1            0     0.000     0.000     0.000          12     0.001     0.002
   0.000 perl
15548     0 sda1            0     0.000     0.000     0.000          12     0.001     0.014
   0.001 perl
15577     0 sda1            0     0.000     0.000     0.000          12     0.001     0.003
   0.000 perl
15519     0 sda1            0     0.000     0.000     0.000          12     0.001     0.005
   0.000 perl
15578     0 sda1            0     0.000     0.000     0.000          12     0.001     0.001
   0.000 perl
15583     0 sda1            0     0.000     0.000     0.000          12     0.001     0.001
   0.000 perl
15547     0 sda1            0     0.000     0.000     0.000          11     0.000     0.002
   0.000 perl
15576     0 sda1            0     0.000     0.000     0.000          11     0.001     0.001
   0.000 perl
15518     0 sda1            0     0.000     0.000     0.000          11     0.000     0.001
   0.000 perl
15354     0 sda1            0     0.000     0.000     0.000          10     0.053     0.053
    0.005 lm_lid.sh
```

The columns are:

PID
: the process ID of the application

UID
: the user ID under which the applications is running

DEV
: the device on which the I/O took place

WRITE_CNT
: the total number of write operations

WRITE_MIN
: the lowest time taken for two consecutive writes (in seconds)

WRITE_MAX
: the greatest time taken for two consecutive writes (in seconds)

WRITE_AVG
: the average time taken for two consecutive writes (in seconds)

READ_CNT
: the total number of read operations

READ_MIN
: the lowest time taken for two consecutive reads (in seconds)

READ_MAX
: the greatest time taken for two consecutive reads (in seconds)

READ_AVG
: the average time taken for two consecutive reads (in seconds)

COMMAND
: the name of the process

In this example, three very obvious applications stand out:

```
  PID   UID DEV     WRITE_CNT WRITE_MIN WRITE_MAX WRITE_AVG    READ_CNT  READ_MIN  READ_MAX
READ_AVG COMMAND
2789  2903 sda1          854     0.000   120.000    39.836           0     0.000     0.000
   0.000 plasma
2573     0 sda1           63     0.033  3600.015   515.226           0     0.000     0.000
   0.000 auditd
2153     0 sda1           26     0.003  3600.029  1290.730           0     0.000     0.000
   0.000 rsyslogd
```

These three applications have a **`WRITE_CNT`** greater than **`0`**, which means that they performed some form of write during the measurement. Of those, **plasma** was the worst offender by a large degree: it performed the most write operations, and of course the average time between writes was the lowest. **Plasma** would therefore be the best candidate to investigate if you were concerned about power-inefficient applications.

Use the **strace** and **ltrace** commands to examine applications more closely by tracing all system calls of the given process ID. In the present example, you could run:

```
strace -p 2789
```

In this example, the output of the **`strace`** contained a repeating pattern every 45 seconds that opened the KDE icon cache file of the user for writing followed by an immediate close of the file again. This led to a necessary physical write to the hard disk as the file metadata (specifically, the modification time) had changed. The final fix was to prevent those unnecessary calls when no updates to the icons had occurred.

2.4. Battery Life Tool Kit

The **Battery Life Tool Kit** (BLTK), is a test suite that simulates and analyzes battery life and performance. BLTK achieves this by performing sets of tasks that simulate specific user groups and reporting on the results. Although developed specifically to test notebook performance, BLTK can also report on the performance of desktop computers when started with the **`-a`**.

BLTK allows you to generate very reproducible workloads that are comparable to real use of a machine. For example, the **office** workload writes a text, corrects things in it, and does the same for a spreadsheet. Running BLTK combined with **PowerTOP** or any of the other auditing or analysis tool allows you to test if the optimizations you performed have an effect when the machine is actively in use instead of only idling. Because you can run the exact same workload multiple times for different settings, you can compare results for different settings.

Install BLTK with the command:

```
yum install bltk
```

Run BLTK with the command:

```
bltk workload options
```

For example, to run the **idle** workload for 120 seconds:

```
bltk -I -T 120
```

The workloads available by default are:

-I, **--idle**
: system is idle, to use as a baseline for comparison with other workloads

-R, **--reader**
: simulates reading documents (by default, with **Firefox**)

-P, **--player**
: simulates watching multimedia files from a CD or DVD drive (by default, with **mplayer**)

-O, **--office**
: simulates editing documents with the **OpenOffice.org** suite

Other options allow you to specify:

-a, **--ac-ignore**
: ignore whether AC power is available (necessary for desktop use)

-T *number_of_seconds*, **--time** *number_of_seconds*
: the time (in seconds) over which to run the test; use this option with the **idle** workload

-F *filename*, **--file** *filename*
: specifies a file to be used by a particular workload, for example, a file for the **player** workload to play instead of accessing the CD or DVD drive

-W *application*, **--prog** *application*
: specifies an application to be used by a particular workload, for example, a browser other than **Firefox** for the **reader** workload

BLTK supports a large number of more specialized options. For details, refer to the **bltk** man page.

BLTK saves the results that it generates in a directory specified in the **/etc/bltk.conf** configuration file — by default, **~/.bltk/*workload*.results.*number*/**. For example, the **~/.bltk/reader.results.002/** directory holds the results of the third test with the **reader** workload (the first test is not numbered). The results are spread across several text files. To condense these results into a format that is easy to read, run:

```
bltk_report path_to_results_directory
```

The results now appear in a text file named **Report** in the results directory. To view the results in a terminal emulator instead, use the **-o** option:

```
bltk_report -o path_to_results_directory
```

2.5. Tuned and ktune

Tuned is a daemon that monitors the use of system components and dynamically tunes system settings based on that monitoring information. Dynamic tuning accounts for the way that various system components are used differently throughout the uptime for any given system. For example, the hard drive is used heavily during startup and login, but is barely used later when a user might mainly work with applications like OpenOffice or email clients. Similarly, the CPU and network devices are used differently at different times. **Tuned** monitors the activity of these components and reacts to changes in their use.

As a practical example, consider a typical office workstation. Most of the time, the Ethernet network interface will be very inactive. Only a few emails will go in and out every once in a while or some web pages might be loaded. For those kinds of loads, the network interface doesn't have to run at full speed all the time, as it does by default. **Tuned** has a monitoring and tuning plugin for network devices that can detect that low activity and then automatically lower the speed of that interface, typically resulting in lower power usage. If activity on the interface increases drastically for a longer period of time, for example because a DVD image is being downloaded or an email with a large attachment is opened, **tuned** detects this and sets the interface speed to maximum to offer the best performance while the activity level is so high. This principle is used for the other plugins for CPU and hard disks as well.

Network devices are not configured to behave this way by default because speed changes can take several seconds to take effect and therefore directly and visibly impact the user experience. Similar considerations apply for the CPU and hard drive tuning plugins. When a hard drive has been spun down, it can take several seconds for it to spin up again which results in an observed lack of responsiveness of the system during that period. The latency side effect is smallest for the CPU plugin, but it is still at least measurable, though hardly noticeable by a user.

Alongside of **tuned** we now also offer **ktune**. **Ktune** was originally a framework and service to optimize the performance of a machine for a specific use cases. Since then, **ktune** has improved to such a degree that we now use it as the static part of our general tuning framework. It is mainly used in the different predefined profiles described in *Section 2.5.2, “Tuned-adm”*.

Install the *tuned* package and its associated **systemtap** scripts with the command:

```
yum install tuned
```

Installing the *tuned* package also sets up a sample configuration file at **/etc/tuned.conf** and activates the default profile.

Start **tuned** by running:

```
service tuned start
```

To start **tuned** every time the machine boots, run:

```
chkconfig tuned on
```

Tuned itself has additional options that you can use when you run it manually. The available options are:

-d, **--daemon**
start **tuned** as a daemon instead of in the foreground.

-c, **--conffile**
use a configuration file with the specified name and path, for example, **--conffile=/etc/tuned2.conf**. The default is **/etc/tuned.conf**.

-D, **--debug**
use the highest level of logging.

2.5.1. The tuned.conf file

The **tuned.conf** file contains configuration settings for **tuned**. By default, it is located at **/etc/tuned.conf**, but you can specify a different name and location by starting **tuned** with the **--conffile** option.

The config file must always contain a **[main]** section that defines the general parameters for **tuned**. The file then contains a section for each plugin.

The **[main]** section contains the following options:

interval
the interval at which **tuned** should monitor and tune the system, in seconds. The default value is **10**.

verbose
specifies whether output should be verbose. The default value is **False**.

logging
specifies the minimum priority of messages to be logged. In descending order, allowable values are: **critical**, **error**, **warning**, **info**, and **debug**. The default value is **info**.

logging_disable
specifies the maximum priority of messages to be logged; any messages with this priority or lower will not be logged. In descending order, allowable values are: **critical**, **error**, **warning**, **info**, and **debug**. The value **notset** disables this option.

Each plugin has its own section, specified with the name of the plugin in square brackets; for example: **[CPUTuning]**. Each plugin can have its own options, but the following apply to all plugins:

enabled
specifies whether the plugin is enabled or not. The default value is **True**.

verbose
specifies whether output should be verbose. If not set for this plugin, the value is inherited from **[main]**.

logging
specifies the minimum priority of messages to be logged. If not set for this plugin, the value is inherited from **[main]**.

A sample config file follows:

```
[main]
interval=10
pidfile=/var/run/tuned.pid
logging=info
logging_disable=notset

# Disk monitoring section

[DiskMonitor]
enabled=True
logging=debug

# Disk tuning section

[DiskTuning]
enabled=True
hdparm=False
alpm=False
logging=debug

# Net monitoring section

[NetMonitor]
enabled=True
logging=debug

# Net tuning section

[NetTuning]
enabled=True
logging=debug

# CPU monitoring section

[CPUMonitor]
# Enabled or disable the plugin. Default is True. Any other value
# disables it.
enabled=True

# CPU tuning section

[CPUTuning]
# Enabled or disable the plugin. Default is True. Any other value
# disables it.
enabled=True
```

The CPU tuning plugin recently added support for the Super Hybrid Engine (SHE), as found on some Asus EEEPCs. If you are using this engine, you can enable SHE support by adding **`eeefsb=True`** into the **`[CPUTuning]`** section of **`tuned.conf`**. It will automatically downclock your FSB and save power while on idle, returning to normal speed when the load increases significantly. This functionality is enabled by default in the **`laptop-battery-powersave`** profile.

2.5.2. Tuned-adm

Often, a detailed audit and analysis of a system can be very time consuming and might not be worth the few extra watts you might be able to save by doing so. Previously, the only alternative was simply to use the defaults. Therefore, Fedora 15 includes separate profiles for specific use cases as an alternative between those two extremes, together with the **tuned-adm** tool that allows you to switch between these profiles easily at the command line. Fedora 15 includes a number of predefined profiles for typical use cases that you can simply select and activate with the **`tuned-adm`** command, but you can also create, modify or delete profiles yourself.

To list all available profiles and identify the current active profile, run:

```
tuned-adm list
```

To only display the currently active profile, run:

```
tuned-adm active
```

To switch to one of the available profiles, run:

```
tuned-adm profile profile_name
```

for example:

```
tuned-adm profile server-powersave
```

To disable all tuning:

```
tuned-adm off
```

When you first install **tuned**, the **default** profile will be active. Fedora 15 also includes the following predefined profiles:

default
: the default power-saving profile. It has the lowest impact on power saving of the available profiles and only enables CPU and disk plugins of **tuned**.

desktop-powersave
: a power-saving profile directed at desktop systems. Enables ALPM power saving for SATA host adapters (refer to *Section 3.6, "Aggressive Link Power Management"*) as well as the CPU, Ethernet, and disk plugins of **tuned**.

server-powersave
: a power-saving profile directed at server systems. Enables ALPM powersaving for SATA host adapters and activates the CPU and disk plugins of **tuned**.

laptop-ac-powersave
: a medium-impact power-saving profile directed at laptops running on AC. Enables ALPM powersaving for SATA host adapters, Wi-Fi power saving, as well as the CPU, Ethernet, and disk plugins of **tuned**.

laptop-battery-powersave
: a high-impact power-saving profile directed at laptops running on battery. It activates all power saving mechanisms from the previous profiles, plus it enables the multi-core power-savings scheduler for low wakeup systems and makes sure that the ondemand governor is active and that AC97 audio power-saving is enabled. On compatible Asus EEEPCs, it enables SHE support in the CPU tuning plugin. You can use this profile to save the maximum amount of power on any kind of system, not only laptops on battery power. The trade-off for this profile is a noticeable impact on performance, specifically latency of disk and network I/O.

spindown-disk
: a power-saving profile for machines with classic HDDs to maximize spindown time. It disables the **tuned** power savings mechanism, disables USB autosuspend, disables Bluetooth, enables Wi-Fi power saving, disables logs syncing, increases disk write-back time, and lowers disk swappiness. All partitions are remounted with **noatime**.

throughput-performance
: a server profile for typical throughput performance tuning. It disables **tuned** and **ktune** power saving mechanisms, enables **sysctl** settings that improve the throughput performance of your disk and network I/O, and switches to the **deadline scheduler**.

latency-performance
: a server profile for typical latency performance tuning. it disables **tuned** and **ktune** power saving mechanisms and enables **sysctl** settings that improve the latency performance of your network I/O.

enterprise-storage
: a server profile directed at enterprise-class storage, maximizing I/O throughput. It activates the same settings as the **`throughput-performance`** profile, multiplies readahead settings, and disables barriers on non-root and non-boot partitions.

All the profiles are stored in separate subdirectories under **`/etc/tune-profiles`**. So **`/etc/tune-profiles/desktop-powersave`** contains all the necessary files and settings for that profile. Each of these directories contains up to four files:

`tuned.conf`
: the configuration for the tuned service to be active for this profile.

`sysctl.ktune`
: the **sysctl** settings used by **ktune**. The format is identical to the **`/etc/sysconfig/sysctl`** file (refer to the sysctl and sysctl.conf man pages).

`ktune.sysconfig`
: the configuration file of ktune itself, typically **`/etc/sysconfig/ktune`**.

`ktune.sh`
: an **init**-style shell script used by the **ktune** service which can run specific commands during system startup to tune the system.

The easiest way to start a new profile is to copy an existing one. The **`laptop-battery-powersave`** profile contains a very rich set of tunings already and is therefore a useful starting point. Simply copy the whole directory to the new profile name like this:

```
cp -a /etc/tune-profiles/laptop-battery-powersave/ /etc/tune-profiles/myprofile
```

Modify any of the files in the new profile to match your personal requirements.

2.6. UPower

In Fedora 11 **DeviceKit-power** assumed the power management functions that were part of **HAL** and some of the functions that were part of **GNOME Power Manager** in previous releases of Fedora (refer also to *Section 2.7, "GNOME Power Manager"*. In Fedora 13, **DeviceKit-power** was renamed to **UPower**. **UPower** provides a daemon, an API, and a set of command-line tools. Each power source on the system is represented as a device, whether it is a physical device or not. For example, a laptop battery and an AC power source are both represented as devices.

You can access the command-line tools with the **`upower`** command and the following options:

`--enumerate`, **`-e`**
: displays an object path for each power devices on the system, for example:

```
/org/freedesktop/UPower/devices/line_power_AC
```

```
/org/freedesktop/UPower/devices/battery_BAT0
```

`--dump`, **`-d`**
displays the parameters for all power devices on the system.

`--wakeups`, **`-w`**
displays the CPU wakeups on the system.

`--monitor`, **`-m`**
monitors the system for changes to power devices, for example, the connection or disconnection of a source of AC power, or the depletion of a battery. Press **`Ctrl`**+**`C`** to stop monitoring the system.

`--monitor-detail`
monitors the system for changes to power devices, for example, the connection or disconnection of a source of AC power, or the depletion of a battery. The **`--monitor-detail`** option presents more detail than the **`--monitor`** option. Press **`Ctrl`**+**`C`** to stop monitoring the system.

`--show-info` ***`object_path`***, **`-i`** ***`object_path`***
displays all information available for a particular object path. For example, to obtain information about a battery on your system represented by the object path `/org/freedesktop/UPower/devices/battery_BAT0`, run:

```
upower -i /org/freedesktop/UPower/devices/battery_BAT0
```

2.7. GNOME Power Manager

GNOME Power Manager is a daemon that is installed as part of the GNOME desktop. Much of the power-management functionality that **GNOME Power Manager** provided in early releases of Fedora became part of **DeviceKit-power** in Fedora 11, renamed to **UPower** in Fedora 13 (refer to *Section 2.6, "UPower"*. However, **GNOME Power Manager** remains a front end for that functionality. Through an applet in the system tray, **GNOME Power Manager** notifies you of changes in your system's power status; for example, a change from battery to AC power. It also reports battery status, and warns you when battery power is low.

GNOME Power Manager also allows you to configure some basic power management settings. To access these settings, click the **GNOME Power Manager** icon in the system tray, then click **Preferences**

Use the **On AC power** and **On battery power** ComboBoxes to specify how much time must pass before an inactive system goes to sleep. You can also choose a standard behavior for a system with a critically low battery. For example, by default, **GNOME Power Manager** makes a system hibernate when its battery level reaches a critically low level. On the bottom of this dialog, you can set behaviors for the physical power button and suspend button on your system.

2.8. Other means for auditing

Fedora 15 offers quite a few more tools with which to perform system auditing and analysis. Most of them can be used as a supplementary sources of information in case you want to verify what you have discovered already or in case you need more in-depth information on certain parts. Many of these tools are used for performance tuning as well. They include:

vmstat

vmstat gives you detailed information about processes, memory, paging, block I/O, traps, and CPU activity. Use it to take a closer look at what the system overall does and where it is busy.

iostat

iostat is similar to **vmstat**, but only for I/O on block devices. It also provides more verbose output and statistics.

blktrace

blktrace is a very detailed block I/O trace program. It breaks down information to single blocks associated with applications. It is very useful in combination with **diskdevstat**.

Core Infrastructure and Mechanics

3.1. CPU Idle States

CPUs with the x86 architecture support various states in which parts of the CPU are deactivated or run at lower performance settings. These states, known as *C-states*, allow systems to save power by partially deactivating CPUs that are not in use. C-states are numbered from C0 upwards, with higher numbers representing decreased CPU functionality and greater power saving. C-States of a given number are broadly similar across processors, although the exact details of the specific feature sets of the state may vary between processor families. C-States 0–3 are defined as follows:

C0
: the operating or running state. In this state, the CPU is working and not idle at all.

C1, Halt
: a state where the processor is not executing any instructions but is typically not in a lower power state. The CPU can continue processing with practically no delay. All processors offering C-States need to support this state. Pentium 4 processors support an enhanced C1 state called C1E that actually is a state for lower power consumption.

C2, Stop-Clock
: a state where the the clock is frozen for this processor but it keeps the complete state for its registers and caches, so after starting the clock again it can immediately start processing again. This is an optional state.

C3, Sleep
: a state where the processor really goes to sleep and doesn't need to keep it's cache up to date. Waking up from this state takes considerably longer than from C2 due to this. Again this is an optional state.

Recent Intel CPUs with the "Nehalem" microarchitecture feature a new C-State, C6, which can reduce the voltage supply of a CPU to zero, but typically reduces power consumption by between 80% and 90%. The kernel in Fedora 15 includes optimizations for this new C-State.

3.2. Using `CPUfreq` Governors

One of the most effective ways to reduce power consumption and heat output on your system is to use CPUfreq. CPUfreq — also referred to as CPU speed scaling — allows the clock speed of the processor to be adjusted on the fly. This enables the system to run at a reduced clock speed to save power. The rules for shifting frequencies, whether to a faster or slower clock speed, and when to shift frequencies, are defined by the CPUfreq governor.

The governor defines the power characteristics of the system CPU, which in turn affects CPU performance. Each governor has its own unique behavior, purpose, and suitability in terms of workload. This section describes how to choose and configure a CPUfreq governor, the characteristics of each governor, and what kind of workload each governor is suitable for.

The main concerns surrounding power management are:

- Heat reduction for servers
- Extending battery life for laptops

As a rule, lowering the power consumption of a specific component (or of the system as a whole) will lead to lower heat and naturally, performance. As such, you should thoroughly study and test the decrease in performance afforded by any configurations you make, especially for mission-critical systems.

The following sections explain how optimal hardware performance benefits your system in terms of energy consumption.

3.2.1. CPUfreq Governor Types

This section lists and describes the different types of CPUfreq governors available in Fedora 15.

cpufreq_performance

The Performance governor forces the CPU to use the highest possible clock frequency. This frequency will be statically set, and will not change. As such, this particular governor offers *no power saving benefit*. It is only suitable for hours of heavy workload, and even then only during times wherein the CPU is rarely (or never) idle.

cpufreq_powersave

By contrast, the Powersave governor forces the CPU to use the lowest possible clock frequency. This frequency will be statically set, and will not change. As such, this particular governor offers maximum power savings, but at the cost of the *lowest CPU performance*.

The term "powersave" can sometimes be deceiving, though, since (in principle) a slow CPU on full load consumes more power than a fast CPU that is not loaded. As such, while it may be advisable to set the CPU to use the Powersave governor during times of expected low activity, any unexpected high loads during that time can cause the system to actually consume more power.

The Powersave governor is, in simple terms, more of a "speed limiter" for the CPU than a "power saver". It is most useful in systems and environments where overheating can be a problem.

cpufreq_ondemand

The Ondemand governor is a dynamic governor that allows the CPU to achieve maximum clock frequency when system load is high, and also minimum clock frequency when the system is idle. While this allows the system to adjust power consumption accordingly with respect to system load, it does so at the expense of *latency between frequency switching*. As such, latency can offset any performance/power saving benefits offered by the Ondemand governor if the system switches between idle and heavy workloads too often.

For most systems, the Ondemand governor can provide the best compromise between heat emission, power consumption, performance, and manageability. When the system is only busy at specific times of the day, the Ondemand governor will automatically switch between maximum and minimum frequency depending on the load without any further intervention.

cpufreq_userspace

The Userspace governor allows userspace programs (or any process running as root) to set the frequency. This governor is normally used in conjunction with the **cpuspeed** daemon. Of all the governors, Userspace is the most customizable; and depending on how it is configured, it can offer the best balance between performance and consumption for your system.

cpufreq_conservative

Like the Ondemand governor, the Conservative governor also adjusts the clock frequency according to usage (like the Ondemand governor). However, while the Ondemand governor does so in a more aggressive manner (that is from maximum to minimum and back), the Conservative governor switches between frequencies more gradually.

This means that the Conservative governor will adjust to a clock frequency that it deems fitting for the load, rather than simply choosing between maximum and minimum. While this can possibly provide significant savings in power consumption, it does so at an ever *greater latency* than the Ondemand governor.

Note

You can enable a governor using **cron** jobs. This allows you to automatically set specific governors during specific times of the day. As such, you can specify a low-frequency governor during idle times (for example after work hours) and return to a higher-frequency governor during hours of heavy workload.

For instructions on how to enable a specific governor, refer to *Procedure 3.2, "Enabling a CPUfreq Governor"* in *Section 3.2.2, "CPUfreq Setup"*.

3.2.2. CPUfreq Setup

Before selecting and configuring a CPUfreq governor, you need to add the appropriate CPUfreq driver first.

Procedure 3.1. How to Add a CPUfreq Driver

1. Use the following command to view which CPUfreq drivers are available for your system:

```
ls /lib/modules/[kernel version]/kernel/arch/[architecture]/kernel/cpu/cpufreq/
```

2. Use **modprobe** to add the appropriate CPUfreq driver.

```
modprobe [CPUfreq driver]
```

When using the above command, be sure to remove the **.ko** filename suffix.

Important

When choosing an appropriate CPUfreq driver, always choose **acpi-cpufreq** over **p4-clockmod**. While using the **p4-clockmod** driver reduces the clock frequency of a CPU, it does not reduce the voltage. **acpi-cpufreq**, on the other hand, reduces voltage along with CPU clock frequency, allowing less power consumption and heat output for each unit reduction in performance.

3. Once the CPUfreq driver is set up, you can view which governor the system is currently using with:

```
cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor
```

You can also view which governors are available for use for a specific CPU using:

```
cat /sys/devices/system/cpu/[cpu ID]/cpufreq/scaling_available_governors
```

Some CPUfreq governors may not be available for you to use. In this case, use **modprobe** to add the necessary kernel modules that enable the specific CPUfreq governor you wish to use. These kernel modules are available in **/lib/modules/*[kernel version]*/kernel/drivers/cpufreq/**.

Procedure 3.2. Enabling a CPUfreq Governor

1. If a specific governor is not listed as available for your CPU, use **modprobe** to enable the governor you wish to use. For example, if the **ondemand** governor is not available for your CPU, use the following command:

```
modprobe cpufreq_ondemand
```

2. Once a governor is listed as available for your CPU, you can enable it using:

```
echo [governor] > /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor
```

3.2.3. Tuning CPUfreq Policy and Speed

Once you've chosen an appropriate CPUfreq governor, you can further tune the speed of each CPU using the tunables found in **/sys/devices/system/cpu/*[cpu ID]*/cpufreq/**. These tunables are:

- **cpuinfo_min_freq** — Shows the CPU's available minimum operating frequency (in KHz).

- **cpuinfo_max_freq** — Shows the CPU's available maximum operating frequency (in KHz).

- **scaling_driver** — Shows what CPUfreq driver is used to set the frequency on this CPU.
- **scaling_available_governors** — Shows the CPUfreq governors available in this kernel. If you wish to use a CPUfreq governor that is not listed in this file, refer to *Procedure 3.2, "Enabling a CPUfreq Governor"* in *Section 3.2.2, "CPUfreq Setup"* for instructions on how to do so.
- **scaling_governor** — Shows what CPUfreq governor is currently in use. To use a different governor, simply use **echo *[governor]* > /sys/devices/system/cpu/*[cpu ID]*/cpufreq/scaling_governor** (refer to *Procedure 3.2, "Enabling a CPUfreq Governor"* in *Section 3.2.2, "CPUfreq Setup"* for more information).
- **cpuinfo_cur_freq** — Shows the current speed of the CPU (in KHz).
- **scaling_available_frequencies** — Lists available frequencies for the CPU, in KHz.
- **scaling_min_freq** and **scaling_max_freq** — Sets the *policy limits* of the CPU, in KHz.

  

  Important

  When setting policy limits, you should set **scaling_max_freq** before **scaling_min_freq**.

- **affected_cpus** — Lists CPUs that require frequency coordination software.
- **scaling_setspeed** — Used to change the clock speed of the CPU, in KHz. You can only set a speed within the policy limits of the CPU (as per **scaling_min_freq** and **scaling_max_freq**).

To view the current value of each tunable, use **cat *[tunable]***. For example, to view the current speed of cpu0 (in KHz), use:

```
cat /sys/devices/system/cpu/cpu0/cpufreq/cpuinfo_cur_freq
```

To change the value of each tunable, use **echo *[value]* > /sys/devices/system/cpu/*[cpu ID]*/cpufreq/*[tunable]***. For example, to set the minimum clock speed of cpu0 to 360 KHz, use:

```
echo 360000 > /sys/devices/system/cpu/cpu0/cpufreq/scaling_min_freq
```

3.3. Suspend and Resume

When a system is suspended, the kernel calls on drivers to store their states and then unloads them. When the system is resumed, it reloads these drivers, which attempt to reprogram their devices. The drivers' ability to accomplish this task determines whether the system can be resumed successfully.

Video drivers are particularly problematic in this regard, because the *Advanced Configuration and Power Interface* (ACPI) specification does not require system firmware to be able to reprogram video hardware. Therefore, unless video drivers are able to program hardware from a completely uninitialized state, they may prevent the system from resuming.

Fedora 15 includes greater support for new graphics chipsets, which ensures that suspend and resume will work on a greater number of platforms.

3.4. Tickless Kernel

Previously, the Linux kernel periodically interrupted each CPU on a system at a predetermined frequency — 100 Hz, 250 Hz, or 1000 Hz, depending on the platform. The kernel queried the CPU about the processes that it was executing, and used the results for process accounting and load balancing. Known as the *timer tick*, the kernel performed this interrupt regardless of the power state of the CPU. Therefore, even an idle CPU was responding to up to 1000 of these requests every second. On systems that implemented power saving measures for idle CPUs, the timer tick prevented the CPU from remaining idle long enough for the system to benefit from these power savings.

The kernel in Fedora 15 runs *tickless*: that is, it replaces the old periodic timer interrupts with on-demand interrupts. Therefore, idle CPUs are allowed to remain idle until a new task is queued for processing, and CPUs that have entered lower power states can remain in these states longer.

3.5. Active-State Power Management

Active-State Power Management (ASPM) saves power in the *Peripheral Component Interconnect Express* (PCI Express or PCIe) subsystem by setting a lower power state for PCIe links when the devices to which they connect are not in use. ASPM controls the power state at both ends of the link, and saves power in the link even when the device at the end of the link is in a fully powered-on state.

When ASPM is enabled, device latency increases because of the time required to transition the link between different power states. ASPM has three policies to determine power states:

default
: sets PCIe link power states according to the defaults specified by the firmware on the system (for example, BIOS). This is the default state for ASPM.

powersave
: sets ASPM to save power wherever possible, regardless of the cost to performance.

performance
: disables ASPM to allow PCIe links to operate with maximum performance.

ASPM policies are set in **`/sys/module/pcie_aspm/parameters/policy`**, but can also be specified at boot time with the *`pcie_aspm`* kernel parameter, where **`pcie_aspm=off`** disables ASPM and **`pcie_aspm=force`** enables ASPM, even on devices that do not support ASPM.

Warning — `pcie_aspm=force` can cause systems to stop responding

If **`pcie_aspm=force`** is set, hardware that does not support ASPM can cause the system to stop responding. Before setting **`pcie_aspm=force`**, ensure that all PCIe hardware on the system supports ASPM.

3.6. Aggressive Link Power Management

Aggressive Link Power Management (ALPM) is a power-saving technique that helps the disk save power by setting a SATA link to the disk to a low-power setting during idle time (that is when there is no I/O). ALPM automatically sets the SATA link back to an active power state once I/O requests are queued to that link.

Power savings introduced by ALPM come at the expense of disk latency. As such, you should only use ALPM if you expect the system to experience long periods of idle I/O time.

ALPM is only available on SATA controllers that use the *Advanced Host Controller Interface* (AHCI). For more information about AHCI, refer to *http://www.intel.com/technology/serialata/ahci.htm*.

When available, ALPM is enabled by default. ALPM has three modes:

min_power

This mode sets the link to its lowest power state (SLUMBER) when there is no I/O on the disk. This mode is useful for times when an extended period of idle time is expected.

medium_power

This mode sets the link to the second lowest power state (PARTIAL) when there is no I/O on the disk. This mode is designed to allow transitions in link power states (for example during times of intermittent heavy I/O and idle I/O) with as small impact on performance as possible.

`medium_power` mode allows the link to transition between PARTIAL and fully-powered (that is "ACTIVE") states, depending on the load. Note that it is not possible to transition a link directly from PARTIAL to SLUMBER and back; in this case, either power state cannot transition to the other without transitioning through the ACTIVE state first.

max_performance

ALPM is disabled; the link does not enter any low-power state when there is no I/O on the disk.

To check whether your SATA host adapters actually support ALPM you can check if the file **`/sys/class/scsi_host/host*/link_power_management_policy`** exists. To change the settings simply write the values described in this section to these files or display the files to check for the current setting.

Important — some settings disable hot plugging

Setting ALPM to **min_power** or **medium_power** will automatically disable the "Hot Plug" feature.

3.7. Relatime Drive Access Optimization

The POSIX standard requires that operating systems maintain file system metadata that records when each file was last accessed. This timestamp is called `atime`, and maintaining it requires a constant series of write operations to storage. These writes keep storage devices and their links busy and powered up. Since few applications make use of the `atime` data, this storage device activity wastes power. Significantly, the write to storage would occur even if the file was not read from storage, but from cache. For some time, the Linux kernel has supported a **noatime** option for **mount** and would not write `atime` data to file systems mounted with this option. However, simply turning off this feature is problematic because some applications rely on `atime` data and will fail if it is not available.

The kernel used in Fedora 15 supports another alternative — **relatime**. **Relatime** maintains `atime` data, but not for each time that a file is accessed. With this option enabled, `atime` data is written to the disk only if the file has been modified since the `atime` data was last updated (`mtime`), or if the file was last accessed more than a certain length of time ago (by default, one day).

By default, all filesystems are now mounted with **relatime** enabled. To suppress this feature across an entire system, use the boot parameter *default_relatime=0*. If **relatime** is enabled on a system by default, you can suppress it for any particular file system by mounting that file system with the option **norelatime**. Finally, to vary the default length of time before which the system will update a file's `atime` data, use the *relatime_interval=* boot parameter, specifying the period in seconds. The default value is **86400**.

3.8. Power Capping

Fedora 15 supports the power capping features found in recent hardware, such as HP *Dynamic Power Capping* (DPC), and Intel Node Manager (NM) technology. Power capping allows administrators to limit the power consumed by servers, but it also allows managers to plan data centers more efficiently, because the risk of overloading existing power supplies is greatly diminished. Managers can place more servers within the same physical footprint and have confidence that if server power consumption is capped, the demand for power during heavy load will not exceed the power available.

HP Dynamic Power Capping

Dynamic Power Capping is a feature available on select ProLiant and BladeSystem servers that allows system administrators to cap the power consumption of a server or a group of servers. The cap is a definitive limit that the server will not exceed, regardless of its current workload. The cap has no effect until the server reaches its power consumption limit. At that point, a management processor adjusts CPU P-states and clock throttling to limit the power consumed.

Dynamic Power Capping modifies CPU behavior independently of the operating system, however, HP's *integrated Lights-Out 2* (iLO2) firmware allows operating systems access to the management processor and therefore applications in user space can query the management processor. The kernel used in Fedora 15 includes a driver for HP iLO and iLO2 firmware, which allows programs to query management processors at **/dev/hpilo/d*X*ccb*N***. The kernel also includes an extension of the `hwmon sysfs` interface to support power capping features, and a `hwmon` driver for ACPI 4.0 power

meters that use the `sysfs` interface. Together, these features allow the operating system and user-space tools to read the value configured for the power cap, together with the current power usage of the system.

For further details of HP Dynamic Power Capping, refer to *HP Power Capping and HP Dynamic Power Capping for ProLiant Servers*, available from *http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01549455/c01549455.pdf*

Intel Node Manager

Intel Node Manager imposes a power cap on systems, using processor P-states and T-states to limit CPU performance and therefore power consumption. By setting a power management policy, administrators can configure systems to consume less power during times when system loads are low, for example, at night or on weekends.

Intel Node Manager adjusts CPU performance using *Operating System-directed configuration and Power Management* (OSPM) through the standard *Advanced Configuration and Power Interface*. When Intel Node Manager notifies the OSPM driver of changes to T-states, the driver makes corresponding changes to processor P-states. Similarly, when Intel Node Manager notifies the OSPM driver of changes to P-states, the driver changes T-states accordingly. These changes happen automatically and require no further input from the operating system. Administrators configure and monitor Intel Node Manager with *Intel Data Center Manager* (DCM) software.

For further details of Intel Node Manager, refer to *Node Manager — A Dynamic Approach To Managing Power In The Data Center*, available from *http://communities.intel.com/docs/DOC-4766*

3.9. Enhanced Graphics Power Management

Fedora 15 saves power on graphics and display devices by eliminating several sources of unnecessary consumption.

LVDS reclocking

Low-voltage differential signaling (LVDS) is a system for carrying electronic signals over copper wire. One significant application of the system is to transmit pixel information to *liquid crystal display* (LCD) screens in notebook computers. All displays have a *refresh rate* — the rate at which they receive fresh data from a graphics controller and redraw the image on the screen. Typically, the screen receives fresh data sixty times per second (a frequency of 60 Hz). When a screen and a graphics controller are linked by LVDS, the LVDS system uses power on every refresh cycle. When idle, the refresh rate of many LCD screens can be dropped to 30 Hz without any noticeable effect (unlike *cathode ray tube* (CRT) monitors, where a decrease in refresh rate produces a characteristic flicker). The driver for Intel graphics adapters built into the kernel used in Fedora 15 performs this *down-clocking* automatically, and saves around 0.5 W when the screen is idle.

Enabling memory self-refresh

Synchronous dynamic random access memory (SDRAM) — as used for video memory in graphics adapters — is recharged thousands of times per second so that individual memory cells retain the data that is stored in them. Apart from its main function of managing data as it flows in and out of memory, the memory controller is normally responsible for initiating these refresh cycles. However, SDRAM also has a low-power *self-refresh* mode. In this mode, the memory uses an internal timer to generate its own refresh cycles, which allows the system to shut down the memory controller without endangering data currently held in memory. The kernel used in Fedora 15 can trigger memory self-refresh in Intel graphics adapters when they are idle, which saves around 0.8 W.

GPU clock reduction

Typical graphical processing units (GPUs) contain internal clocks that govern various parts of their internal circuitry. The kernel used in Fedora 15 can reduce the frequency of some of the internal clocks in Intel and ATI GPUs. Reducing the number of cycles that GPU components perform in a given time saves the power that they would have consumed in the cycles that they did not have to perform. The kernel automatically reduces the speed of these clocks when the GPU is idle, and increases it when GPU activity increases. Reducing GPU clock cycles can save up to 5 W.

GPU power-down

The Intel and ATI graphics drivers in Fedora 15 can detect when no monitor is attached to an adapter and therefore shut down the GPU completely. This feature is especially significant for servers which do not have monitors attached to them regularly.

3.10. RFKill

Many computer systems contain radio transmitters, including Wi-Fi, Bluetooth, and 3G devices. These devices consume power, which is wasted when the device is not in use.

RFKill is a subsystem in the Linux kernel that provides an interface through which radio transmitters in a computer system can be queried, activated, and deactivated. When transmitters are deactivated, they can be placed in a state where software can reactive them (a *soft block*) or where software cannot reactive them (a *hard block*).

The RFKill core provides the application programming interface (API) for the subsystem. Kernel drivers that have been designed to support RFkill use this API to register with the kernel, and include methods for enabling and disabling the device. Additionally, the RFKill core provides notifications that user applications can interpret and ways for user applications to query transmitter states.

The RFKill interface is located at **`/dev/rfkill`**, which contains the current state of all radio transmitters on the system. Each device has its current RFKill state registered in **`sysfs`**. Additionally, RFKill issues *uevents* for each change of state in an RFKill-enabled device.

Rfkill is a command-line tool with which you can query and change RFKill-enabled devices on the system. To obtain the tool, install the *rfkill* package.

Use the command **`rfkill list`** to obtain a list of devices, each of which has an *index number* associated with it, starting at **`0`**. You can use this index number to tell **rfkill** to block or unblock a device, for example:

```
rfkill block 0
```

blocks the first RFKill-enabled device on the system.

You can also use **rfkill** to block certain categories of devices, or all RFKill-enabled devices. For example:

```
rfkill block wifi
```

blocks all Wi-Fi devices on the system. To block all RFKill-enabled devices, run:

```
rfkill block all
```

To unblock devices, run **`rfkill unblock`** instead of **`rfkill block`**. To obtain a full list of device categories that **rfkill** can block, run **`rfkill help`**

3.11. Optimizations in User Space

Reducing the amount of work performed by system hardware is fundamental to saving power. Therefore, although the changes described in *Chapter 3, Core Infrastructure and Mechanics* permit the system to operate in various states of reduced power consumption, applications in user space that request unnecessary work from system hardware prevent the hardware from entering those states. During the development of Fedora 15, audits were undertaken in the following areas to reduce unnecessary demands on hardware:

Reduced wakeups

Fedora 15 uses a *tickless kernel* (refer to *Section 3.4, "Tickless Kernel"*), which allows the CPUs to remain in deeper idle states longer. However, the *timer tick* is not the only source of excessive CPU wakeups, and function calls from applications can also prevent the CPU from entering or remaining in idle states.

Reduced storage and network IO

Input or output (IO) to storage devices and network interfaces forces devices to consume power. In storage and network devices that feature reduced power states when idle (for example, ALPM or ASPM), this traffic can prevent the device from entering or remaining in an idle state, and can prevent hard drives from spinning down when not in use. Excessive and unnecessary demands on storage have been minimized in several applications. In particular, those demands that prevented hard drives from spinning down.

Initscript audit

Services that start automatically whether required or not have great potential to waste system resources. Services instead should default to "off" or "on demand" wherever possible. For example, the **BlueZ** service that enables Bluetooth support previously ran automatically when the system started, whether Bluetooth hardware was present or not. The **BlueZ** initscript now checks that Bluetooth hardware is present on the system before starting the service.

Use Cases

This chapter describes two types of use case to illustrate the analysis and configuration methods described elsewhere in this guide. The first example considers typical servers and the second is a typical laptop.

4.1. Example — Server

A typical standard server nowadays comes with basically all of the necessary hardware features supported in Fedora 15. The first thing to take into consideration is the kinds of workloads for which the server will mainly be used. Based on this information you can decide which components can be optimized for power savings.

Regardless of the type of server, graphics performance is generally not required. Therefore, GPU power savings can be left turned on.

Webserver

A webserver needs network and disk I/O. Depending on the external connection speed 100 Mbit/s might be enough. If the machine serves mostly static pages, CPU performance might not be very important. Power-management choices might therefore include:

- no disk or network plugins for **tuned**.
- ALPM turned on.
- `ondemand` governor turned on.
- network card limited to 100 Mbit/s.

Compute server

A compute server mainly needs CPU. Power management choices might include:

- depending on the jobs and where data storage happens, disk or network plugins for **tuned**; or for batch-mode systems, fully active **tuned**.
- depending on utilization, perhaps the `performance` governor.

Mailserver

A mailserver needs mostly disk I/O and CPU. Power management choices might include:

- `ondemand` governor turned on, because the last few percent of CPU performance are not important.
- no disk or network plugins for **tuned**.
- network speed should not be limited, because mail is often internal and can therefore benefit from a 1 Gbit/s or 10 Gbit/s link.

Fileserver

Fileserver requirements are similar to those of a mailserver, but depending on the protocol used, might require more CPU performance. Typically, Samba-based servers require more CPU than NFS, and NFS typically requires more than iSCSI. Even so, you should be able to use the `ondemand` governor.

Directory server

A directory server typically has lower requirements for disk I/O, especially if equipped with enough RAM. Network latency is important although network I/O less so. You might consider latency network tuning with a lower link speed, but you should test this carefully for your particular network.

4.2. Example — Laptop

One other very common place where power management and savings can really make a difference are laptops. As laptops by design normally already use drastically less energy than workstations or servers the potential for absolute savings are less than for other machines. When in battery mode, though, any saving can help to get a few more minutes of battery life out of a laptop. Although this section focuses on laptops in battery mode, but you certainly can still use some or all of those tunings while running on AC power as well.

Savings for single components usually make a bigger relative difference on laptops than they do on workstations. For example, a 1 Gbit/s network interface running at 100 Mbits/s saves around 3–4 watts. For a typical server with a total power consumption of around 400 watts, this saving is approximately 1 %. On a laptop with a total power consumption of around 40 watts, the power saving on just this one component amounts to 10 % of the total.

Specific power-saving optimizations on a typical laptop include:

- Configure the system BIOS to disable all hardware that you do not use. For example, parallel or serial ports, card readers, webcams, Wi-Fi, and Bluetooth just to name a few possible candidates.
- Dim the display in darker environments where you do not need full illumination to read the screen comfortably. On the GNOME desktop, use **Applications**+**System Tools** → **System Settings**, then select **Hardware** → **Power**. On the KDE desktop, use **Kickoff Application Launcher**+**Computer**+**System Settings**+**Advanced** → **Power Management**. Alternatively, enter **gnome-power-manager** or **xbacklight** at the command line or use the function keys on your laptop.
- Use the `laptop-battery-powersave` profile of **tuned-adm** to enable a whole set of power-saving mechanisms. Note that performance and latency for the hard drive and network interface are impacted.

Additionally (or alternatively) you can perform many small adjustments to various system settings:

- use the `ondemand` governor (enabled by default in Fedora 15)
- enable laptop mode (part of the `laptop-battery-powersave` profile):

```
echo 5 > /proc/sys/vm/laptop_mode
```

- increase flush time to disk (part of the `laptop-battery-powersave` profile):

```
echo 1500 > /proc/sys/vm/dirty_writeback_centisecs
```

- disable nmi watchdog (part of the `laptop-battery-powersave` profile):

```
echo 0 > /proc/sys/kernel/nmi_watchdog
```

- enable AC97 audio power-saving (enabled by default in Fedora 15):

```
echo Y > /sys/module/snd_ac97_codec/parameters/power_save
```

- enable multi-core power-saving (part of the `laptop-battery-powersave` profile):

```
echo Y > /sys/module/snd_ac97_codec/parameters/power_save
```

- enable USB auto-suspend:

```
for i in /sys/bus/usb/devices/*/power/autosuspend; do echo 1 > $i; done
```

  Note that USB auto-suspend does not work correctly with all USB devices.

- enable minimum power setting for ALPM (part of the `laptop-battery-powersave` profile):

```
echo min_power > /sys/class/scsi_host/host*/link_power_management_policy
```

- mount filesystem using relatime (default in Fedora 15):

```
mount -o remount,relatime mountpoint
```

- activate best power saving mode for hard drives (part of the `laptop-battery-powersave` profile):

```
hdparm -B 1 -S 200 /dev/sd*
```

- reduce screen brightness to **`50`** or less, for example:

```
xbacklight -set 50
```

- activate DPMS for screen idle:

```
xset +dpms; xset dpms 0 0 300
```

- reduce Wi-Fi power levels (part of the `laptop-battery-powersave` profile):

```
for i in /sys/bus/pci/devices/*/power_level ; do echo 5 > $i ; done
```

- deactivate Wi-Fi:

```
echo 1 > /sys/bus/pci/devices/*/rf_kill
```

- limit wired network to 100 Mbit/s (part of the `laptop-battery-powersave` profile):

```
ethtool -s eth0 advertise 0x0F
```

Appendix A. Tips for Developers

Every good programming textbook covers problems with memory allocation and the performance of specific functions. As you develop your software, be aware of issues that might increase power consumption on the systems on which the software runs. Although these considerations do not affect every line of code, you can optimize your code in areas which are frequent bottlenecks for performance.

Some techniques that are often problematic include:

- using threads.
- unnecessary CPU wake-ups and not using wake-ups efficiently. If you must wake up, do everything at once (race to idle) and as quickly as possible.
- using `[f]sync()` unnecessarily.
- unnecessary active polling or using short, regular timeouts. (React to events instead).
- not using wake-ups efficiently.
- inefficient disk access. Use large buffers to avoid frequent disk access. Write one large block at a time.
- inefficient use of timers. Group timers across applications (or even across systems) if possible.
- excessive I/O, power consumption, or memory usage (including memory leaks)
- performing unnecessary computation.

The following sections examine some of these areas in greater detail.

A.1. Using Threads

It is widely believed that using threads makes applications perform better and faster, but this is not true in every case.

Python

Python uses the Global Lock Interpreter[1], so threading is profitable only for larger I/O operations. **Unladen-swallow** [2] is a faster implementation of Python with which you might be able to optimize your code.

Perl

Perl threads were originally created for applications running on systems without forking (such as systems with 32-bit Windows operating systems). In Perl threads, the data is copied for every single thread (Copy On Write). Data is not shared by default, because users should be able to define the level of data sharing. For data sharing the **threads::shared** module has to be included. However, data is not only then copied (Copy On Write), but the module also creates tied variables for the data, which takes even more time and is even slower. [3]

[1] *http://docs.python.org/c-api/init.html#thread-state-and-the-global-interpreter-lock*
[2] *http://code.google.com/p/unladen-swallow/*
[3] *http://www.perlmonks.org/?node_id=288022*

C

C threads share the same memory, each thread has its own stack, and the kernel does not have to create new file descriptors and allocate new memory space. C can really use the support of more CPUs for more threads. Therefore, to maximize the performance of your threads, use a low-level language like C or C++. If you use a scripting language, consider writing a C binding. Use profilers to identify poorly performing parts of your code. [4]

A.2. Wake-ups

Many applications scan configuration files for changes. In many cases, the scan is performed at a fixed interval, for example, every minute. This can be a problem, because it forces a disk to wake up from spindowns. The best solution is to find a good interval, a good checking mechanism, or to check for changes with **inotify** and react to events. **Inotify** can check variety of changes on a file or a directory.

For example:

```
#include <stdio.h>
#include <stdlib.h>
#include <sys/time.h>
#include <sys/types.h>
#include <sys/inotify.h>
#include <unistd.h>

int main(int argc, char *argv[]) {
  int fd;
  int wd;
  int retval;
  struct timeval tv;

  fd = inotify_init();

  /* checking modification of a file - writing into */
  wd = inotify_add_watch(fd, "./myConfig", IN_MODIFY);
  if (wd < 0) {
    printf("inotify cannot be used\n");
    /* switch back to previous checking */
  }

  fd_set rfds;
  FD_ZERO(&rfds);
  FD_SET(fd, &rfds);
  tv.tv_sec = 5;
  tv.tv_usec = 0;
  retval = select(fd + 1, &rfds, NULL, NULL, &tv);
  if (retval == -1)
    perror("select()");
  else if (retval) {
    printf("file was modified\n");
  }
  else
    printf("timeout\n");

  return EXIT_SUCCESS;
}
```

The advantage of this approach is the variety of checks that you can perform.

[4] *http://people.redhat.com/drepper/lt2009.pdf*

The main limitation is that only a limited number of watches are available on a system. The number can be obtained from **`/proc/sys/fs/inotify/max_user_watches`** and although it can be changed, this is not recommended. Furthermore, in case **inotify** fails, the code has to fall back to a different check method, which usually means many occurrences of **`#if #define`** in the source code.

For more information on **inotify**, refer to the inotify man page.

A.3. Fsync

`Fsync` is known as an I/O expensive operation, but this is is not completely true. For example, refer to Theodore Ts'o's article *Don't fear the fsync!* [5] and the accompanying discussion.

Previously, **Firefox** called the **sqlite** library each time the user clicked on a link to go to a new page. **Sqlite** called `fsync` and because of the file system settings (mainly ext3 with data-ordered mode), there was a long latency when nothing happened. This could take a long time (30 seconds or more) if another process was copying a large file at the same time.

However, in other cases, where `fsync` wasn't used at all, problems emerged with the switch to the ext4 file system. Ext3 was set to data-ordered mode, which flushed memory every few seconds and saved it to a disk. But with ext4 and laptop_mode, the interval between saves was longer and data might get lost when the system was unexpectedly switched off. Now ext4 is patched, but we must still consider the design of our applications carefully, and use `fsync` as appropriate.

The following simple example of reading and writing into a configuration file shows how a backup of a file can be made or how data can be lost:

```
/* open and read configuration file e.g. ~/.kde/myconfig */
fd = open("./kde/myconfig", O_WRONLY|O_TRUNC|O_CREAT);
read(myconfig);
...
write(fd, bufferOfNewData, sizeof(bufferOfNewData));
close(fd);
```

A better approach would be:

```
open("/.kde/myconfig", O_WRONLY|O_TRUNC|O_CREAT);
read(myconfig);
...
fd = open("/.kde/myconfig.suffix", O_WRONLY|O_TRUNC|O_CREAT);
write(fd, bufferOfNewData, sizeof(bufferOfNewData));
fsync; /* paranoia - optional */
...
close(fd);
/* do_copy("/.kde/myconfig", "/.kde/myconfig~"); */ /* paranoia - optional */
rename("/.kde/myconfig.suffix", "/.kde/myconfig");
```

[5] *http://thunk.org/tytso/blog/2009/03/15/dont-fear-the-fsync/*

Appendix B. Revision History

Revision 1.0-1 Wed Sep 28 2011 **Jack Reed** *jreed@redhat.com*

Incorporated patch by Jaroslav Skarvada

Revision 0.1 Thu Jul 29 2010 **Rüdiger Landmann** *r.landmann@redhat.com*

Bring upstream