

The GPL Compliance Engineering Guide

v3.5 - 2010-04-02

Armijn Hemel <armijn@loohuis-consulting.nl>

Copyright © 2009-2010 Loohuis Consulting. Verbatim copying and distribution of this entire article is permitted in any medium, provided this notice is preserved.

Table of Contents

Introduction.....	4
The consumer electronics business.....	4
How a product is developed.....	5
Violations.....	5
Technical analysis of a device.....	6
Initial network scan.....	6
How to perform a network scan.....	6
Results of a network scan.....	6
Value of using network scans.....	7
Other network tricks.....	7
Firmware analysis.....	7
Embedded design 101.....	8
Boot sequence and boot loaders.....	8
Compression techniques.....	8
File systems.....	9
squashfs.....	10
ext2/ext3/ext4.....	11
cramfs.....	11
jffs2.....	11
yaffs2.....	12
Executable files.....	12
Compilation 101.....	12
Executable formats.....	13
Tools.....	13
File analysis tools.....	14
hexdump.....	14
file.....	14
strings.....	14
grep.....	15
md5sum/sha1sum/sha256sum/sha512sum.....	15
Tools for unpacking files and archives.....	15
bzip2/bzcat.....	16
gzip/zcat.....	16
unzip.....	16
lzma.....	16
unrar.....	17
cabextract.....	17
unshield.....	17
rpmdevtools/rpm2cpio.....	17
Other tools.....	17
binutils.....	18
ldd.....	18
editor.....	18
Physical access.....	18
Serial console.....	18
Attaching a serial cable to a router.....	18

Accessing the serial port.....	22
JTAG.....	22
What violations to look for.....	23
Linux kernel modules.....	23
busybox.....	23
C libraries.....	24
Toolchain.....	24
Bootloaders.....	25
Physical compliance.....	26
Compliance engineering on Microsoft Windows.....	26
Common violations.....	26
Tools.....	27
Zipped executables.....	27
Cabinet files.....	27
MSI files.....	27
Wine.....	27
Other tools.....	27
Cygwin compliance engineering.....	28
Experiences.....	28
Appendix A: GPL checklist.....	29
Appendix B: Reporting and fixing license violations.....	29
Reporting a violation.....	29
Handling a violation report	30
Preventing a violation	30
Copyright note	31
Appendix C: Commercial compliance engineering.....	31

Introduction

This is a guide explaining how to find license violations in embedded devices. This guide shows how to discover problems by analysis of network scans, extracting information from a firmware and physically altering hardware.

Before we can dive into the technical details, it is worth taking a look at the business processes of the consumer electronics industry, where most violations are found.

WARNING: Some things described in this guide might not be allowed in your jurisdiction due to local legislation. Please consult a lawyer to see what is permitted. This is not legal advice.

The consumer electronics business

The consumer electronics business is challenging. The business itself is very high volume and has very low margin. Competition in this market is very fierce. The shelf life of a typical device is short: 1 to 1.5 years. Most of the sales of a new product happen during the first 3 months the device is on the market. The consumers mostly look at functionality and price. Time to market, marketing, price and emotional attachment to a particular brand are what drives the market.

Making it to the shops a few days later than a competitor's product (which ironically often comes from the same factories) could mean the difference between having a profit or turning a loss. Raising the price, even by small amounts like 10 cents per device, could mean the same.

Compliance engineering and checking for licensing issues tends to endanger profit. First of all, it delays the release. Proper compliance engineering could take a few days (depending on the device), any questions regarding sources have to go back to the factory, sources have to be shipped, and so on. Often the factory won't or can't release all sources (because they bought it too) and it could take many months before the device is compliant. Arriving a few months later than the competition will mean you lost the race. Companies often also don't get more than one or two test samples, which they cannot afford to lend out to a compliance engineer when they need to test functionality.

The second reason is that compliance engineering in general is not cheap and the costs of it have to be split per device. A price of EUR 1200 for checking a device is reasonable, given the hourly rates for a commercial embedded Linux systems engineer. Still, for companies in this market this is a lot of money, especially if you keep in mind that many companies have so called test runs of hardware to test demand in the market. A test run is done with as few as 200 devices. If a product is selling, additional shipments are ordered at the factory. For GPL compliance the amount of devices does not matter, since distribution is distribution, but EUR 1200 divided by 200 means a sharp raise in the price of a device.

Companies often have to make a choice: ship non-compliant software and risk a court case or face a huge loss resulting from missed sales. Some people have hinted that a court case is unlikely to happen and is probably a lot cheaper than the alternative. Various organisations,

like the gpl-violations.org project and SFLC have started pushing for compliance a lot more in the past years, so this argument is likely to become invalid soon.

How a product is developed

Products are often not developed by the company that has its name on the box. There are few Western companies selling devices in large quantities to end consumers that do their own development. Even these companies that do are unlikely to do all the work themselves.

There are often quite a few companies involved in the development of a product. The Western companies buy their devices in Asia, most often from a Taiwanese, Chinese or sometimes a Korean company. In some cases a custom casing is developed for the product, but more often a generic casing is adapted with the company logo printed on the casing. The manual and packaging are also adapted to taste (company logos, contact information, etcetera) and everything is shipped to the West. The Western companies do distribution, marketing, end user support, rebates, and so on.

The company where the devices are produced use a board design with a SDK, which they get from another upstream vendor, often the chip vendor. There can be additional layers in between. The engineers at the Taiwanese company, or any of the other layers, sometimes add some extra code, or make other changes using the SDK. The extra code might contain kernel drivers for various hardware components in the device, such as wireless network cards, or software firewalling modules.

These changes may be fully, partially or not at all integrated into the source archive from the SDK. If the sources are not or partially integrated the result is that the sources distributed as the "GPL sources" are not complete.

Violations

License violations come in all kinds of forms, ranging from forgetting to add a copy of the license text to no source, no license text and no policy of handling source code requests. License violations are not limited to just GPL and LGPL. Nearly every device that runs Linux also has a whole range of other subtle violations of MIT, BSD and other licenses.

There are also plenty of GPL license violations on devices that don't run Linux. There are for example devices that run a very basic proprietary operating system, but also include some GPL licensed code, which is linked into one big binary blob along with the rest of the operating system.

This document will mainly focus on GPL and LGPL license compliance engineering on Linux systems, with a small section dedicated on analysing common data formats on Microsoft Windows.

Technical analysis of a device

A technical analysis is the technical part of the GPL compliance engineering process. The goal of a technical analysis is to determine if there is GPL or LGPL licensed software on a device. A technical analysis can be performed in several ways. Often there is device, firmware, source tarball (or any combination thereof) that you are asked to check for compliance. Depending on the situation, a lot of work could be required to discover whether GPL violations exist, or to make sure there are none. This can range from dissecting a firmware and go as far as physical modification of a device to log in via a serial port onto the device, or beyond. This section summarizes my tools of choice to do this. It is far from complete and I am very certain I do not find all violations. Still, it is more than what most people at companies are able (or willing) to find. The more violations you catch, the more pressure we can put on a company to adopt better internal processes to prevent violations from happening at all in the future.

Initial network scan

If the device is networked, it is a good idea to start a scan from the network. Many operating systems have slight differences in the networking stack in how they respond to certain packets. Using heuristics, where a lot of specially crafted packets are sent to the device it is possible to determine what a device runs. Scanning tools like nmap have a fingerprinting option, which is fairly accurate, though not fool proof.

How to perform a network scan

The most feature rich network scanning tool on Linux and other Unix-like operating systems is nmap. A typical commandline invocation for fingerprinting a device would look like this:

```
# nmap -P0 -O <ip address> -p 1-65535
```

This command does not first try to ping the device on the network (-P0), which is a considerable speedup, since many devices do not respond to pings. The fingerprinting option (-O) needs root privileges to work.

Results of a network scan

The output of an invocation of the nmap command could look like this:

```
# nmap -P0 -O 10.0.1.1
```

```
Starting Nmap 4.20 ( http://insecure.org ) at 2007-09-16 01:16 CEST
Interesting ports on gateway.local (10.0.1.1):
Not shown: 1692 closed ports
PORT      STATE SERVICE
22/tcp    open  ssh
25/tcp    open  smtp
53/tcp    open  domain
139/tcp   open  netbios-ssn
445/tcp   open  microsoft-ds
MAC Address: 00:00:00:00:00:00
```

Device type: general purpose
Running: FreeBSD 6.X
OS details: FreeBSD 6.1-RELEASE through 6.2-BETA3 (x86)
Uptime: 36.216 days (since Fri Aug 10 20:06:29 2007)
Network Distance: 1 hop

OS detection performed. Please report any incorrect results at
<http://insecure.org/nmap/submit/> .
Nmap finished: 1 IP address (1 host up) scanned in 21.304 seconds

The result is a list of open TCP ports on the device with an indication which service these ports are normally assigned to and the name and version of the operating system that nmap thinks the device is running (in this case it correctly identified a FreeBSD machine).

Scanning for open TCP ports can reveal important information. Sometimes there is a debug port, or telnet port which you can connect to to gain shell access to the device itself and inspect it while it is running. Banner strings for programs (webserver, UPnP server, FTP server, etcetera) can also help in determining what is running on a particular device.

Value of using network scans

While a network scan is powerful it should not be regarded as proof. There are software packages, so called scrubbers, which will hide a lot of these details and make fingerprinting relatively useless. Not many devices deploy such techniques, so for now it can be regarded as a good indication what is running, until all devices use scrubbers. There are sometimes also false positives, where a device seems to run Linux. If you are not sure you should always perform the scan with a different scanorder (with the -r option of nmap) too.

The output from nmap should be regarded as an indication of how likely it is a device is interesting for compliance engineering. It depends on what you are looking for. Just keep in mind that nmap is not always correct (though very often it is).

Other network tricks

There are some other tricks you can use to get some more information. In the web interface of a device you might see version strings of programs hidden in logfiles. Sometimes the web interfaces on devices have security bugs which can lead to access to the file system. This way you can quickly obtain the data on the device.

Firmware analysis

A reliable method of finding GPL licensed code in a device is by grabbing the firmware of the device from the download site or CD and dissecting it to reveal all bits and pieces of what is in the firmware. There is no standard recipe for dissecting firmware, since there are many ways the firmware of a device can be structured. However, the underlying methodologies as outlined in this document can be used for many devices. Before these methodologies are explained there is a short explanation of how devices work and why the design influences the layout of the firmware.

Embedded design 101

There are a few basic design steps you will need to know for proper reverse engineering. These are:

- boot sequence and boot loaders
- file systems
- compression techniques
- executable formats

A good book to read to get some general understanding about embedded Linux is "Building Embedded Linux Systems", published by O'Reilly.

<http://www.oreilly.com/catalog/belinuxsys/>

Boot sequence and boot loaders

When a device starts the CPU executes certain commands to initialize the whole device. One of the things that is hardcoded in the CPU is the memory location of an instruction that should be loaded first. This instruction is often the first instruction of the bootloader. The bootloader is a program which sets up the rest of the system.

The bootloader itself resides on flash memory at a fixed offset. This offset varies greatly between CPUs, boards and vendors.

As an example, a fairly typical layout for the flash chip of a device with the AR7 chipset could look like this:

```
mt d0    0x900a0000, 0x903f0000
mt d1    0x90010000, 0x900a0000
mt d2    0x90000000, 0x90010000
mt d3    0x903f0000, 0x90400000
```

These regions are used by the bootloader. In this particular case the bootloader itself resides on "mtd2" and this is the location that the CPU uses to find the bootloader. The bootloader reads the other locations from nvrAm to find the kernel and the root file system and load and start accordingly.

The offsets between the file systems can often be seen in the firmware. To create the right offsets between different parts of the firmware something called padding is used. This often consists of zeroes, or other padding characters (0xff seems to be popular too). This makes it easy to recognize the different parts, since there will be a lot of these padding characters together (up to several thousand in some cases).

Compression techniques

Sometimes file systems (squashfs, ext2) or a kernel image can be found directly, but they are often in compressed form in the firmware and have to be decompressed first. During start up of the device the boot loader decompresses the kernel and/or file systems in memory, before actually launching the OS.

Common used compression methods are gzip and bzip2, with LZMA and 7z rapidly rising in popularity.

File systems

There are a couple of file systems in use on embedded Linux devices. They can be divided into two categories. The file systems in the first category load the file system from flash and uncompress it into normal memory. The file systems in the second category don't load into memory, but use more flash to reduce wear levelling on the flash memory. Both approaches have their advantages and disadvantages.

Commonly used file systems are:

- squashfs, increasingly with LZMA compression instead of zlib compression
- ext2fs/ext3fs
- cramfs (Compressed ROM File System), both big endian/little endian
- romfs
- jffs2
- yaffs2

Most of these file systems can be unpacked or mounted over loopback on a recent Linux system (like Fedora 11).

The following table summarizes the methods you should use for the most commonly used file systems:

File system	Unpacking method	Alternative unpacking method	Remarks
SquashFS (zlib compression)	unsquashfs	mount over loopback	
SquashFS (LZMA compression)	custom unsquashfs (for example from OpenWrt)	mount over loopback, might require an extra kernel module, depending on the flavour used	various combinations of SquashFS and LZMA are in use
ext2/ext3	mount over loopback	e2tools package	
cramfs	mount over loopback		might require byteswapping with cramfsswap first, depending on the endianness of your machine
romfs	mount over loopback		

jffs2	jffs2dump	copy content to mtd device first, then mount over loopback	
yaffs2	unyaffs		

squashfs

Squashfs is a read only file system for Linux. It is a popular choice in embedded devices. Standard versions of squashfs can be found in a firmware file by looking for the string 'sqsh' (big endian format) or 'hsqs' (little endian format). Other variants of squashfs might have different magic strings and can't be unpacked with the standard tools.

The squashfs file system (with zlib compression) can also be unpacked as a normal user, using "unsquashfs" from the squashfs-tools package:

```
$ unsquashfs -d rootdir -i /path/to/squashfs-image
```

This command unpacks the squashfs image in the directory "rootdir". This method is actually preferable to mounting over loopback, since it won't create device files if you run it as a normal user and prevent you from mistakes later on, such as trying to grep through tty files (which has rather unpleasant side effects).

Unpacking a squashfs file system with LZMA compression is possible in some cases, but not in all cases. The reason for this is that there are quite a few versions of LZMA in use, which are not always compatible. The Squashfs LZMA version at <http://www.squashfs-lzma.org/> for example uses different magic and it can't work with many Squashfs filesystems that are actually used on embedded devices.

It is not possible to detect LZMA compression using the command "file", since the signature is usually not different from an uncompressed squashfs file system. When you try to mount it and it fails, you might see this in dmesg, which is a clear indication another compression technique than zlib has been used:

```
SQUASHFS: Mounting a different endian SQUASHFS filesystem on loop0
SQUASHFS error: zlib_inflate returned unexpected result 0xffffffff, srclength 8192,
avail_in 160, avail_out 8192
SQUASHFS error: sb_bread failed reading block 0x4b0
SQUASHFS error: Unable to read cache block [12bf5c:3d6]
SQUASHFS error: Unable to read inode [12bf5c:3d6]
```

The OpenWrt project builds a version of "unsquashfs" with LZMA support by default (called "unsquashfs-lzma"), since June 2008. With this tool it is possible to extract Squashfs 3.0 filesystems that use LZMA compression. Older versions of Squashfs can't be uncompressed with it. It is expected that this will be possible in newer versions, as soon as Squashfs with LZMA compression is accepted in the mainline kernel.

ext2/ext3/ext4

The default file system on most Linux systems are the ext2, ext3 and ext4 file systems. These can simply be mounted on the majority of systems over loopback. Some kernels don't have support for this file system built in (rarely), or sometimes you have no root access to mount an ext2 file system over loopback. In such cases the e2tools package provides a barebones way to access an ext2 file system from userspace:

```
$ e2ls ramdisk_e1
bin          boot          default      dev          etc          home
image.cfs   lib           lost+found  mnt         proc        root
sbin        sys          tmp         usr         var         web

$ e2ls ramdisk_e1:etc
TZ          fstab          ftpaccess    ftpaccess.default
ftpconversions  ftpmaxnumber  hotplug     inittab
nsswitch.conf  rc.d          samba
```

cramfs

Another popular file system is the cramfs file system. It can be fairly easily recognized by searching for the string "Compressed ROMFS". There are two versions: one for big endian systems (PowerPC, SPARC, big endian MIPS) and little endian systems (x86, little endian MIPS).

Depending on which system you work on these file systems might need to be byteswapped from big endian to little endian, or vice versa if you want to mount it on over loopback on a Linux system. The cramfsswap utility is a tool that can change the endianness of a cramfs file system.

Byte swapping will not always work, since some devices (notably with the bcm63xx chipset) have a patched cramfs implementation, but it is often enough to extract at least the directory hierarchy and names of the files on the device, which will often give you more information about what is actually on the device.

jffs2

The jffs2 file system is special, since it can't be mounted directly over loopback. It first needs to be written to a special device in memory, which can then be mounted as a normal file system. For this some dark kernel voodoo magic is needed.

The jffs2 file system comes in two flavours: little endian and big endian. Big endian file systems can't be mounted on little endian file systems and vice versa. It might be necessary to convert the endianness of the file system with a program such as jffs2dump before you can access its contents.

The mtd-utils package contains all tools necessary to work with flash memory devices. One of the most useful tools is jffs2dump. With jffs2dump you can inspect the structure file systems and change endianness, or dump the contents of the file system.

A rule of thumb is that if you dump the contents of the jffs2 file (using -c) and you get a lot of warnings, but no real data, you should supply one of the options -b (big endian) or -l (little endian), depending on the endianness of your own system.

Mounting over loopback is possible by first writing the contents of the file to a mtd device and then mounting it.

```
modprobe mtdcore
modprobe jffs2
modprobe mtdram
modprobe mtblock
modprobe mtdchar
dd if=/path-to-jffs2-file of=/dev/mtd0
mount -t jffs2 /dev/mtblock0 /tmp/mnt/
```

This should be enough to mount the file system. The default size that the mtd device can hold is 4 MB. Sometimes there are bigger jffs2 file systems than that and you have to supply a size parameter when loading the mtdram module:

```
modprobe mtdram total_size=8192
```

This will create a ramdisk sized 8 megabytes.

yaffs2

A recent file system is yaffs2. While it has so far been spotted on just a few devices, it is expected to be used a lot more in the near future on embedded devices. There is a unyaffs tool, but it will require some fiddling to actually unpack the data.

Executable files

Executable files are usually the "real" programs on a device. There are two types of executable files:

- scripts
- compiled programs

Scripts can be GPL licensed too, but since they tend to be human readable anyway this is often regarded as not having the highest priority.

The focus of GPL compliance engineering is mostly on compiled programs, which have been transformed from a human readable format into a machine readable format by a process called "compilation".

Compilation 101

Compilation is the process of turning a piece of human readable code into a machine executable program. It starts with someone writing a program in a programming language like C or C++. The compiler analyses the program (this is called "parsing") and translates it into object files. The object files are then linked into an executable, or into general purpose libraries, so they can be used by multiple programs.

There are two types of linking. The first one is static linking, where all functionality that is needed is compiled into one standalone binary file. This includes (parts of) the system C library and all other libraries that are needed to make the program run. Static linking is done

at compile time.

Dynamic linking works differently. The linking phase is postponed until the program is actually executed. A program called "dynamic linker" combines the program with the libraries that need to be loaded to make the program run.

License wise there are no differences between the two (an often made mistake), but the reverse engineering process might differ.

Executable formats

There are a few types of executable formats you can find on an embedded device:

- ELF with/without gzip compression, stripped and not stripped
- Binary Flat format (bFLT) with/without gzip compression

The ELF format is the most common format. Most of the time the binaries will be "stripped", which means that all the debugging information has been removed from the file. If you are lucky the binary has not been stripped and all this information will still be there. This gives more clues about what is actually in the file.

A rare form of the ELF format is where the programs are compressed with gzip, after the ELF header. To get to the contents of the file you first have to extract the contents from the file. This is done in the same way as you would extract a file system which has been compressed with gzip.

The ELF format is an industry standard. There are a lot of tools which can be used to inspect ELF binaries from all kinds of platforms. The GNU binutils collection contains a few tools for doing exactly this: readelf and objdump.

One of the interesting sections in the ELF format is the so called 'dynamic section'. In this section the dynamically linked libraries are listed:

```
$ objdump -x <file> | grep NEEDED
```

There is a lot more functionality that readelf and objdump offer, but the bulk of violations are not discovered that way.

Another format is the Binary Flat Format, or bFLT. It is the default on uClinux based systems and not used on normal Linux systems. This format is more space efficient than ELF, but also contains less information which can be used to identify strings inside programs. There are also fewer tools available with which you can inspect the binaries (other than just dumping the strings). As with the ELF format, there is a special variant which uses gzip compression that has to be unpacked first.

Tools

The toolbox of a reverse engineer contains a lot of tools. The toolset can be divided in a few categories:

- file analysis tools

- tools for unpacking files and archives
- other tools

File analysis tools

hexdump

The hexdump utility is a very valuable tool for reverse engineering. It displays the contents of a file, with offsets and ASCII translations, if the '-C' option is used. It outputs via standard output, so you will need a pager, such as 'more' or 'less' to catch its output. Example output would look like this:

```
00012da0  20 64 6f 6e 65 2c 20 62  6f 6f 74 69 6e 67 20 74  | done, booting t|
00012db0  68 65 20 6b 65 72 6e 65  6c 2e 0a 00 1f 8b 08 00  |he kernel.....|
00012dc0  3b b2 2a 44 02 03 ec bd  7d 7c 54 57 b9 36 bc f6  |;. *D....}|TW.6..|
```

With this you can quickly spot interesting text ("kernel") and the gzip header ("1f 8b 08") immediately following it. Padding can be easily be spotted because hexdump "compresses" this information for you by using '*':

```
00007ee0  ff ff ff ff ff ff ff ff  ff ff ff ff ff ff ff ff  |.....|
*
00010000  27 05 19 56 af a1 29 38  44 2a b2 3f 00 0b 11 b8  |'..V..)8D*.?.....|
```

file

The "file" tool quickly lets you determine what a file might contain. It does so by looking at the first so many bytes and comparing that with known signatures from the so called "magic" file, which on a Linux system can usually be found in /usr/share/magic.

```
$ file main-fs
main-fs: Squashfs filesystem, big endian, version 3.0, 9319589 bytes, 1498 inodes,
blocksize: 65536 bytes, created: Fri Aug 10 14:33:39 2007
```

Often firmware will just show up as "data":

```
$ file zImage
zImage: data
```

This is because a filesystem often has a header or other bytes (padding) put in front of it. Using "file" is not a 100% foolproof method. Sometimes a match for a file system or compressed file is found, while in reality there is no file system or compressed file there. Also "file" won't detect every file system. Before you use it, always try to have the latest version of the magic database installed on your system.

strings

The "strings" tool comes in handy when you want to extract readable strings from a binary file. The strings you extract from binaries are often gibberish, but the readable parts you can get out of a binary are often very helpful and contain function names, literal output written by

programs (for example `kprintf()` statements), and so on. These strings, combined with a search engine or knowledge base of known strings, can reveal a lot.

grep

The "grep" tool is great for quickly finding strings in files (even binaries) that can be important. "Copyright" (with and without capitalization), "Free Software", "License", "GPL" and "General Public License" are good strings to search for. If you specify the command line option "-i" your searches will be case insensitive and quite a bit slower. I usually search for "icense", or "opyright", omitting the first character, which may or not be capitalized. It often saves me a few minutes waiting.

Be warned, many file systems contain special device files or symbolic links to /tmp or other parts of your own file system. If you're not careful you might be grepping on your whole computer, or 'grep' might be stuck on a special device file. A good idea is to first filter out the right files with for example "find" and then grep through them.

md5sum/sha1sum/sha256sum/sha512sum

Fingerprinting tools, like md5sum and other tools from the SHA family, come in handy for identifying files. They work by taking the contents of a file and creating a cryptographic checksum. Two files that are identical will have the same checksum.

Since the MD5 and SHA1 algorithms are known to have "collisions" (two files can have the same fingerprint) it is advised to use sha256sum or sha512sum instead.

Tools for unpacking files and archives

Compression/file format	Unpacking tool	Alternative	Remarks
gzip	gunzip	zcat	zcat unpacks to stdout by default and needs to be redirected to a file
bzip2	bunzip2	bzcat	bzcat unpacks to stdout by default and needs to be redirected to a file
ZIP	unzip		
lzma	lzcat	unlzma	lzcat unpacks to stdout by default and needs to be redirected to a file
tar	tar		
cpio	cpio		

Windows executable	7z	cabextract, unshield, WINE	often the quickest way to extract files is by using WINE
RAR	unrar		
7zip	7z		
rpm	rpmdevtools	rpm2cpio	

bzip2/bzcat

Data compressed with bzip2 can be easily found by searching for the string "BZh" inside the firmware image.

gzip/zcat

A gzip header as used in most devices starts with the header "1f 8b 08" (hexadecimal). Using "hexdump -C" these can be easily found. If you look with vi or vim, then these three characters are formatted as "^_<8b>" by vim. Files can be easily unpacked by using zcat and redirecting output to a file:

```
$ zcat infile > outfile
```

unzip

Normal ZIP files can be unpacked using the unzip program. In firmwares ZIP compressed parts normally start with "PK". Some Windows executables can also be unpacked with unzip.

lzma

A technique that is becoming increasingly popular is LZMA compression. Its claim is that it offers better compression than other compression techniques. There is support for LZMA decompression in various bootloaders and it is fairly popular for Squashfs file systems as a replacement for zlib compression.

An example from a file that uses LZMA compression:

```
00000020  4c 69 6e 75 78 20 4b 65 72 6e 65 6c 20 49 6d 61 |Linux Kernel Ima|
00000030  67 65 00 00 00 00 00 00 00 00 00 00 00 00 00 |ge.....|
00000040  5d 00 00 00 02 00 b0 2a 00 00 00 00 00 00 00 6f |].....*.....o|
```

The LZMA compressed file can be recognized by the sequence '5d 00 00', at hex offset 0x040. Unpacking can be done by the 'lzcat' or 'lzma' tools. It does not support offsets, so in this case the first 64 bytes should be removed, after which it can be unpacked:

```
$ lzma -cd infile > outfile
```

Another tool that is convenient is 'lzmainfo', which gives a lot of information about a file compressed with LZMA:

```
$ lzmainfo lzma-file
```

```
lzma-file
Uncompressed size:      3 MB (2797568 bytes)
Dictionary size:        32 MB (2^25 bytes)
Literal context bits (lc): 3
Literal pos bits (lp):  0
Number of pos bits (pb): 2
```

unrar

The RAR archiving format is very popular on Windows, because of its (claimed) superior compression rates. A lot source archives are distributed in this format. To unpack these files on Linux/Unix you should use the "unrar" program. This program is distributed as freeware and there is currently no free software alternative.

cabextract

The "cabextract" utility is a program to extract cabinet (.cab) archives. This is an archive format which is commonly used on Microsoft Windows. Since GPL license violations are not limited to just Unix(-like) platforms, but also can occur on Microsoft Windows this is a useful tool to have. Some ActiveX components, for example to make an IP camera work with Internet Explorer, are distributed as a cabinet archive. Sometimes firmware updates for Linux based devices are also shipped inside a Microsoft Windows executable.

unshield

The "unshield" utility is used to extract InstallShield cabinet files. These files are a variant of the .cab archives that "cabextract" cannot extract. As of the time of writing the officially released version of unshield cannot extract all types of InstallShield files. There are patches available in the source code repository for unshield that add this type of functionality.

rpmdevtools/rpm2cpio

Sometimes source code archives contain files in the RPM format. RPM is the native format for the installer on various Linux distributions. In some source archives for devices the sources are distributed as RPM files. With rpm2cpio the RPM file can be converted to a cpio archive, which can be extracted using cpio. A recent development is rpmdevtools, which allows easy unpacking of RPMs. Since RPM will move to a new internal format (using 7z) this is the preferred way.

Other tools

- binutils
- ldd (part of a C library, like GNU libc)
- any editor that can read in binary files, such as vi or emacs
- base64

binutils

The binutils package contains several useful tools to inspect binaries, such as readelf and nm.

ldd

The ldd tool prints shared libraries for a dynamically linked executable.

editor

A proper editor is used if you want to edit files and extract parts. Alternatively, a tool such as dd can be used.

Physical access

The final part of compliance engineering work is getting physical access to a device. Sometimes the bootloader is not shipped in a firmware update and can only be accessed through a serial console or JTAG. Often a GPL licensed bootloader is used on a device. If you don't perform a check using a serial port, it can easily be missed.

Serial console

Many devices have a serial port, or a serial port can be attached to it without too much effort. A serial port is used during development of the device. The firmware of the device often lets you log in on the device via the serial port when you connect to it through a serial cable, or gives you a root shell on the device directly. This is not always guaranteed to work. In some devices no output is sent to the serial port during booting, or once the device has booted.

Attaching a serial cable to a router

You can log onto the serial port by using a cable, which attaching one side to the serial port on the router and the other port to the PC, either a serial port on the motherboard, or a serial-USB converter.

WARNING: Many routers work on 3.3 Volts, while a serial port on a PC works on 12 Volts. You need a special cable which can shift between the two voltages or you risk blowing up the device.

There are special kits (MAX 232) to make so called "level shifters", that take care of the voltage difference. Old Siemens phone cables also work. Some online shops sell RS 232 shifters:

http://www.sparkfun.com/commerce/product_info.php?products_id=449

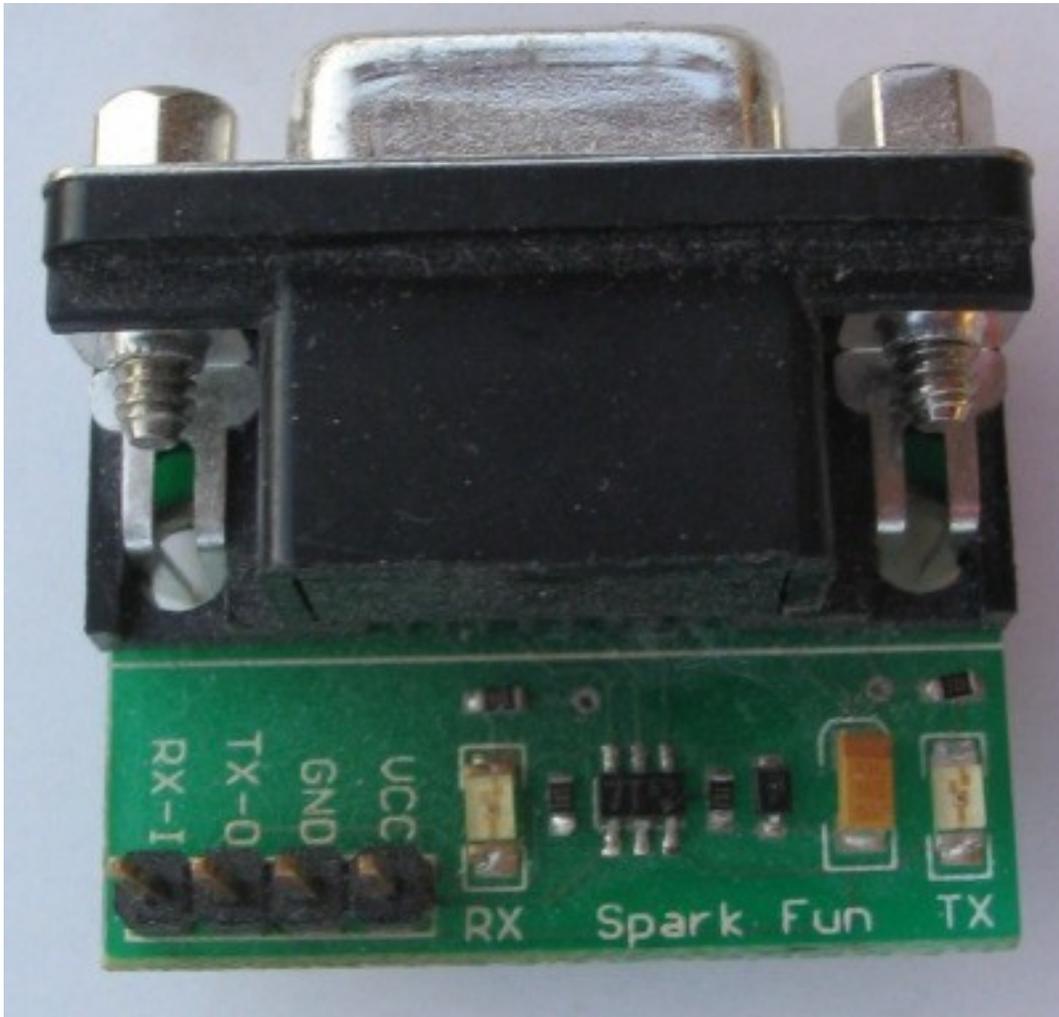


Illustration 1: Pre-made level shifter.

Before you can connect the onboard serial port to a PC you often have to solder header pins onto the solder pads for the serial port. Often there are four or more solder pads next to each other on a device, sometimes even labeled as "serial", "COM1", or equivalent. These header pins can be obtained at any decent electronics store, for a few cents per pin.

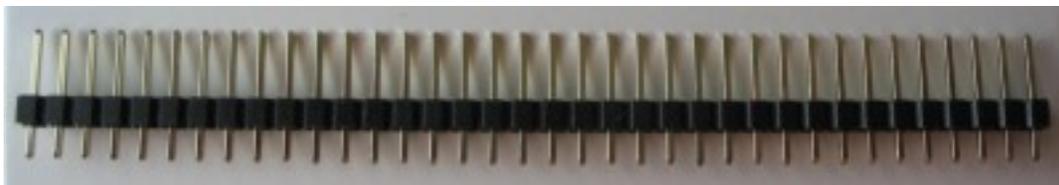


Illustration 2: A row of 36 header pins.

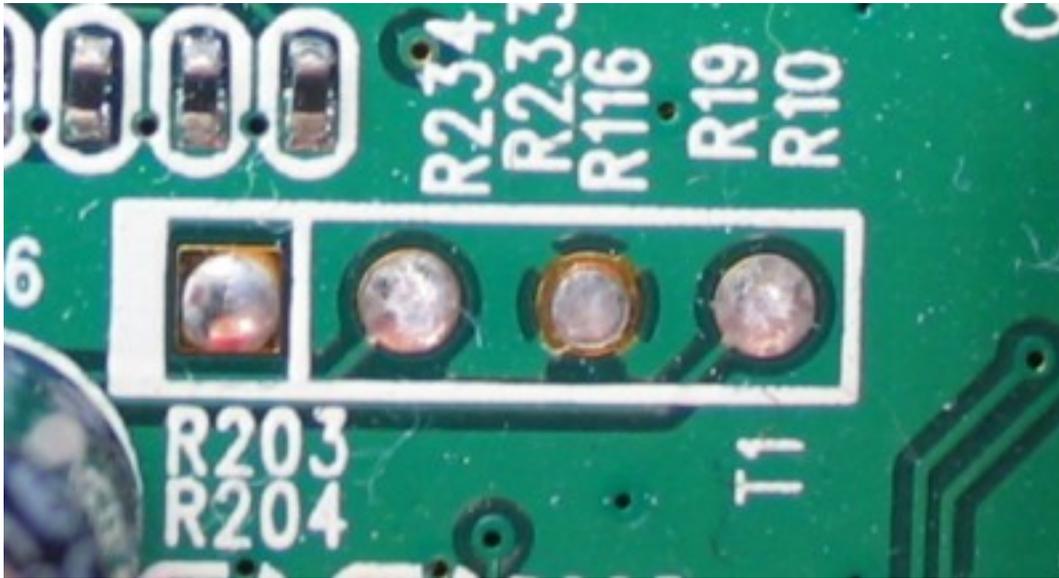


Illustration 3: Solder pads for a serial port on a device, without header pins

Some vendors try to hide these solder pads to make physical access to the device harder. Luckily most vendors simply don't care and in some of the devices you can already find pin headers soldered onto the solder pads.

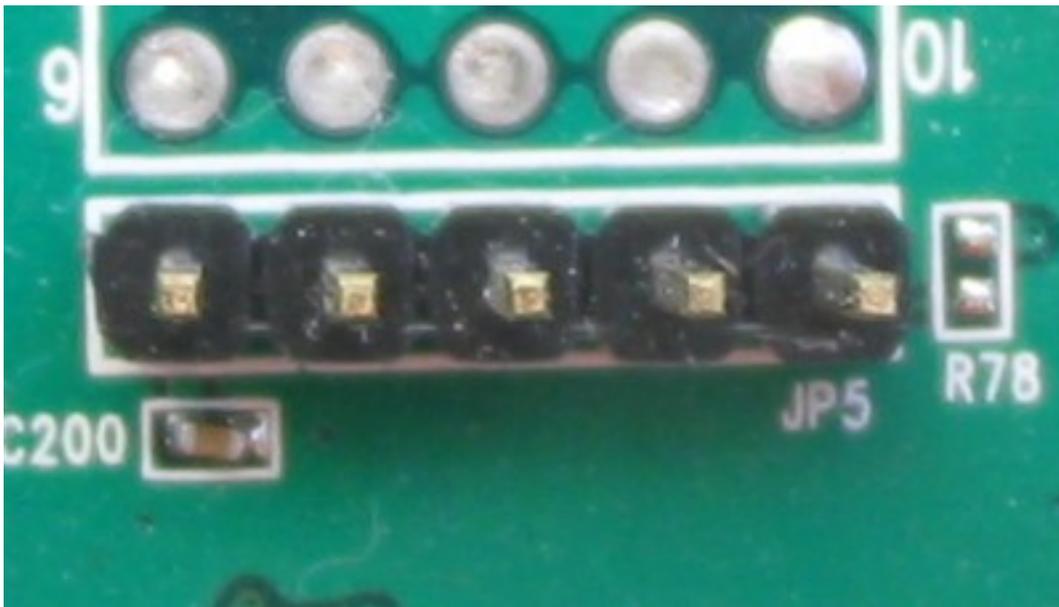


Illustration 4: Serial port with header pins pre-soldered.

Of the solder pads usually only four are needed:

- GND (ground)
- Tx (transmit)
- Rx (receive)
- VCC

These can be mapped to equivalent ports on a serial port on the PC. Different boards have different layouts, often varying between models and revisions.

A serial port can be discovered by using a multimeter to measure the voltage on the pins or solder pads.

- GND: 0 volts
- VCC: 3.3 volts
- Tx/Rx: variable

Many boards use a default layout and people have already made the effort of finding out where the ports on a wide variety of boards are located. The OpenWrt wiki is a great (though somewhat chaotic) resource for this. It is advised to always verify with a multimeter if the information is correct.

The next step is having a proper cable. A cheap solution is to use CD-ROM audio cables. Some of these have connectors that can easily be reused. The connectors look like this:

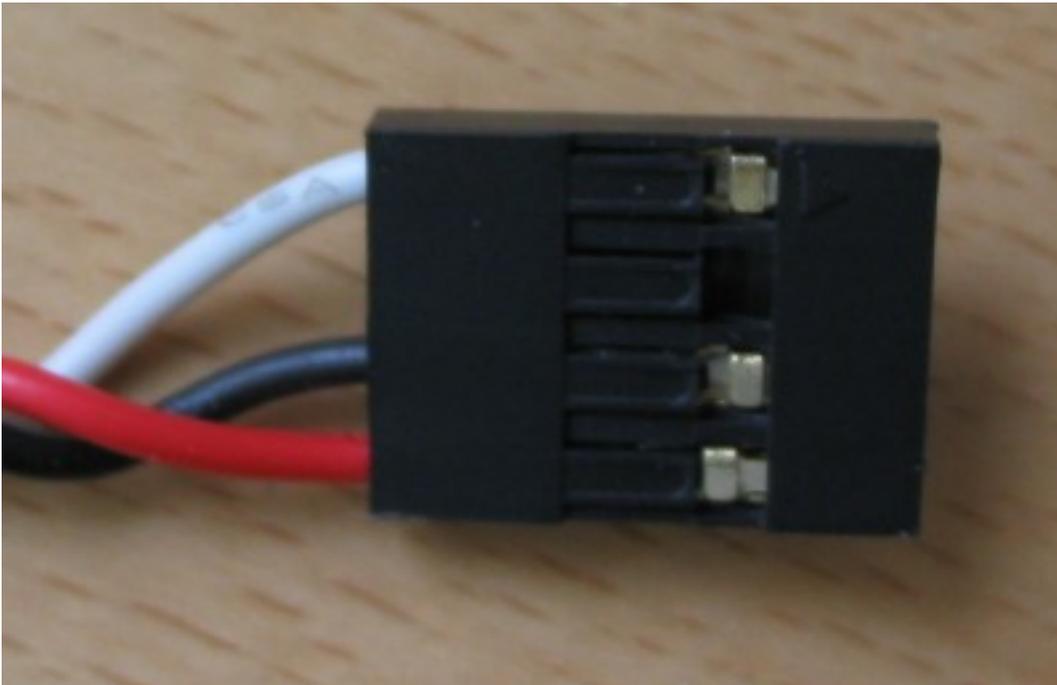


Illustration 5: CD-ROM audio cable.

The black clips that are holding the connectors can easily be lifted, so the cable, plus connector, can be removed by simply pulling the cable. The connectors can then be attached to the pins on the serial port.

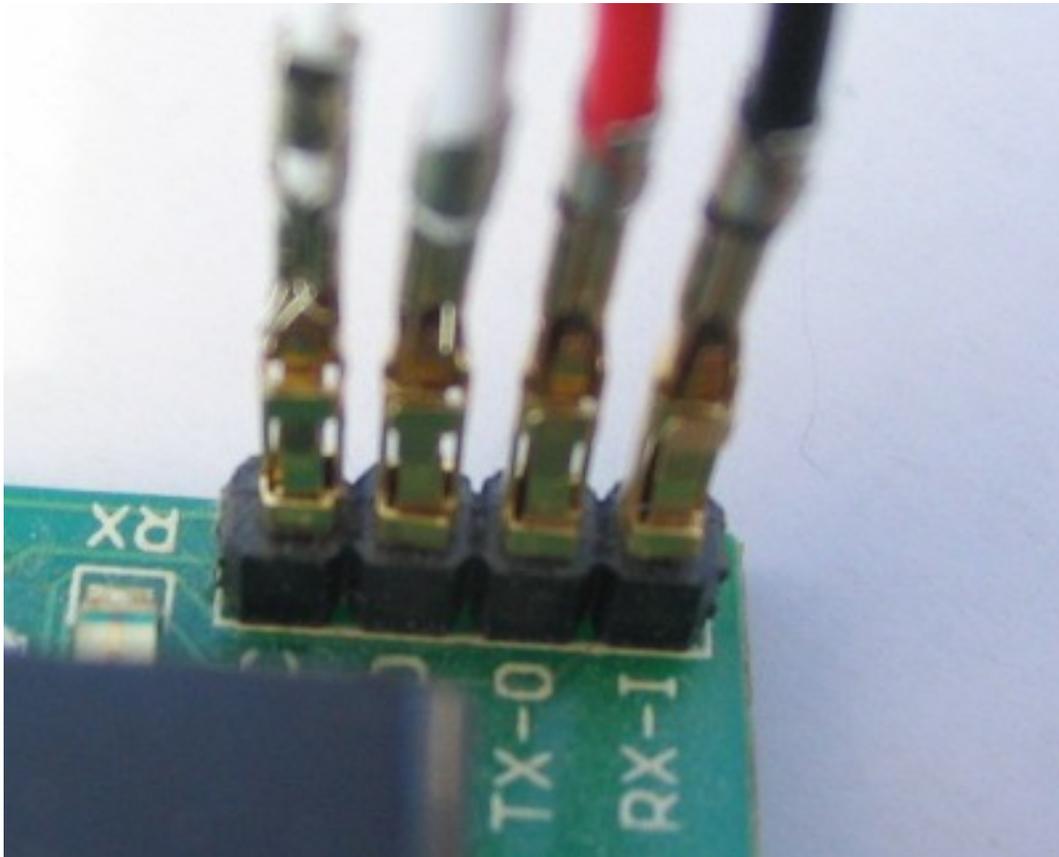


Illustration 6: Modified CD-ROM audio cable attached to header pins.

Accessing the serial port

When the serial cable has been properly attached to the router it can be accessed using a serial communication program. The most popular one on Linux is called 'minicom'. Not all serial ports use the same speed (or 'baud rate'). Popular baud rates are 9600, 38400, 57600 and 115200.

JTAG

Some devices can only be accessed through JTAG.

What violations to look for

A lot of packages are common on most devices. Depending on the package concerned, different techniques will have to be employed to find violations. A few common ones are described below.

Linux kernel modules

One of the grey areas of Linux kernel licensing has been kernel modules. There are a lot of kernel modules which are not licensed as GPL. In the past there has been ongoing discussions whether or not the modules should be GPL licensed or not.

One of the first measurements to clarify the status is the use of a "license" macro in kernel modules:

```
# strings rt2500.o | grep license
license=GPL
```

Modules that have set this macro will have access to more internals of the kernel. Licensing of modules that have this macro set should never be an issue.

With regard to other modules opinions differ. Greg Kroah-Hartman, one of the leading Linux kernel developers, told me in a personal e-mail on 14 October 2007:

[[It's quite simple, me, and my lawyers feel that there is NO way to have a Linux kernel module that is not under the GPLv2. To do so otherwise violates the license of the kernel, and my copyrights. But it's not only me that says this, Novell and IBM have publicly stated this in the past, as well as HP (well, they kind of murmured it, but have said so in person.) Red Hat also states this, as well as a number of key Linux kernel contributors and holders of copyright on the kernel.

The Linux Foundation also issued a statement on closed source drivers and modules on June 23 2008:

http://www.linuxfoundation.org/en/Device_driver_statement

Appendix C of the book "Building Embedded Linux Systems" (1st edition), published by O'Reilly, also has 11 pages dedicated to how kernel developers see the legal status of binary kernel modules. Although the mails are dated (in the time period 1999-2002) and the authors of the mails are not legal professionals, they do provide an insight into the subject.

busybox

Busybox is a program that combines a lot of functionality of programs into one, while leaving out the more advanced features of many of the GNU tools. It is the Swiss army knife of embedded Linux and nearly default on embedded Linux devices. It works by making a symlink from a program to the busybox binary. Depending on as which program it is invoked it will behave differently.

By default not all functionality is built into busybox. At compilation time a configuration (much

like the configuration for the Linux kernel, using a curses based interface) is read to determine which functionality should be built into busybox. The configuration commonly resides in a file called ".config". The configuration file is written after the configuration utility is run.

Options that are enabled are set like this in the configuration file:

```
CONFIG_CAT=y
```

Options that are disabled are set as:

```
# CONFIG_CHGRP is not set
```

Inside the busybox binary you can find hints about which configuration is used, depending on things like the version of busybox that is used. Sometimes the string "Currently defined functions" is followed by a list of functions, which maps more or less directly to the configuration of busybox. In other cases a shell command like the following might be more useful (note: this only works for busybox executables linked with uClibc):

```
$ strings busybox | grep _main | sort
```

This should give you a list with various function names, like:

```
vi_main  
wc_main  
wget_main  
which_main  
yes_main
```

The part before '_main' matches with the name of the applet (executable), which can easily be matched with the busybox configuration. Vice versa this is not always true, since some configuration options are just options to tweak applets, not to build new ones.

In general, the symlinks to the busybox binary and the functions defined in the binary should match the busybox configuration, or else it is a violation of the license.

C libraries

A Linux system is not complete without the so called C library, which contains functionality every program on the system, apart from the Linux kernel itself, is using one way or another. There are two C libraries on Linux that are popular on embedded Linux systems (except Android phones): glibc and uClibc. Another C library that is sometimes (but not often) used is dietlibc. Both glibc and uClibc are LGPL licensed, while dietlibc is GPLv2 licensed. For many embedded devices sources for these libraries are missing, because the C library is often part of the so called toolchain.

Toolchain

An often overlooked part in the compliance process is the toolchain. A toolchain is the combination of a compiler, C library, header files and binutils that can translate programs written by a programmer to something a computer understands.

The compiler parses, checks and translates the source code and generates machine readable code for the platform it was told to generate code for. In most cases, that is the same platform it is running on. So, for example, on my PC I compile a program with the standard compiler that Fedora 11 ships. The output of the compilation process will be a program that can run on my PC. If I would be developing for another platform, based on the MIPS or ARM architecture (or another platform, or another operating system) I would have to instruct my compiler to generate code that will run on that platform, because programs for my Intel x86 based PC will not run on a box that uses a MIPS CPU and runs NetBSD. For this you need a special setup of compiler, plus assembler and linker (found in GNU binutils) that can generate code for a specific platform and a C library to turn it into a working executable. This is not something the standard compilers on standard Linux distributions do by default (note: toolchains are not specific for embedded devices. The combination of compiler, binutils C library and header files on my normal PC is also a toolchain).

The task of building a cross compiler is not trivial and quite tricky to get right (it even gets a lot more fun when you try to cross compile a cross compiler). There are a lot of build environments that make it easy to build a complete development environment for a certain platform, including a proper toolchain. OpenWrt and buildroot are two popular ones, but a lot of vendors have their own build environment, which is shipped as part of a Software Development Kit (SDK). These SDKs, while containing a lot of GPL and LGPL licensed code, are often (partially) included in source distributions in binary form, or not shipped (many vendors have proprietary tools inside the toolchain and don't allow their customers to redistribute the SDK), often resulting in missing sources for the (LGPL/GPL licensed) C library.

Some vendors, such as Broadcom, have adapted the GNU Compiler Collection (GCC) and GNU binutils to take advantage of/use specific characteristics of their CPU. Without these extensions to the compiler you will never be able to create a new program and run it on a machine with code generated with that compiler (the situation might not be as black and white as I put it here, but it makes things definitely a lot harder).

It is an ongoing debate whether or not the toolchain itself should be shipped with a source tarball as part of the obligations described in the GPL. Some people say it should be, since without it it is very difficult and sometimes even impossible to build a new executable for a device without having access to the exact cross compiler that was used for building the software. Other people say that because only the result of the toolchain is distributed, the toolchain does not need to be distributed.

It is beyond any doubt that if a toolchain is available in binary form in the GPL sources for a device and it contains GPL or LGPL code (gcc, binutils, glibc, uClibc or dietlibc) the licenses should be adhered to.

Bootloaders

There are a few GPL licensed bootloaders that are popular in current embedded products. In compliance engineering these are often overlooked.

Bootloader	platforms	comments
------------	-----------	----------

PPCBoot	PowerPC	discontinued, but still used occasionally
ARMboot	ARM	
u-boot	various	
RedBoot	various	originally from eCos, modified GPL license

To find out if these bootloaders are used it is often necessary to access the device through the serial port.

Physical compliance

The physical compliance requirements various licenses have are often overlooked. Compliance engineering is not complete without an inspection of the documentation that is shipped with a device.

The GPL and LGPL licenses require that a copy of the license is shipped with the device, either physically (for example, as part of the manual) or on a documentation CD-ROM. Quite often a device is not shipped with either of them, or just the GPL, even if LGPL licensed code is in used which is the case in nearly all Linux based devices (a notable exception is Android based phones).

Compliance engineering on Microsoft Windows

Most GPL violations we know of are on embedded systems running Linux. There appear to be plenty of violations in programs that run on Microsoft Windows too. The reason that these violations are fairly unknown is that they have never been a focal point for compliance engineering, mostly due to lack of research into this area.

Common violations

A common report is of shareware programs, like CD/DVD burning programs, or music players, that are being distributed in a GPL in compliant way. The 'creators' of those programs tend to be rather immune to requests for the source code and keep happily violating the GPL and LGPL licenses.

Other reported violations are programs using parts of Cygwin, for example in management software for various expensive access points. Other common violations are using the GPL licensed versions of the Qt toolkit or XviD.

An interesting area of research for violations is in ActiveX components that are shipped with for example IP cameras or routers. The ActiveX components are on the device itself and are downloaded by the web browser from the device to get some extra functionality, such as viewing data, or controlling a camera. This is software too and it should also be checked for violations.

Tools

There are a few common archive formats for Windows executables and shared libraries. Which one is used depends on which packaging program was used.

Zipped executables

Quite often files with the '.exe' extension are in fact self extracting executables which have been compressed using ZIP. These can easily be extracted with the 'unzip' program. After unpacking other methods can be used to further investigate the contents.

Cabinet files

A common archiving format for Windows executables is the 'cabinet archive'. A cabinet archive often has the .cab file extension. On Unix systems the "cabextract" and "unshield" tools can be used to extract these files.

MSI files

Another file format that is used a lot is the Microsoft Installer Format, which can be recognized by the .msi file extension. Often you can extract the data from the MSI using the '7z' program. Sometimes this will not work and you will have to try other methods (like the one described next). Extracting a .msi file directly with cabextract will usually get you the file names, but not the content of those files.

After unpacking with 7z you will usually see a lot of that were inside a MSI file, such as resources (pictures, helpfiles) but also shared libraries (DLL) and cabinet archives, which can be extracted as described above.

Wine

A very useful tool to extract data from Windows installers is Wine. During installation data such as archives are written to temporary locations in the file system (C:\windows\temp\). During or after installation these archives or the binaries on the system can be easily copied to another place and analysed using one of the methods described above.

Other tools

On Windows different file formats are used than on Linux and most tools described earlier document to inspect binaries won't work. For example, on Linux the ELF executable format is primarily used, but on Windows the PE executable format is used. Binaries in PE format keep their data in a different form, in such a way that tools like "strings" are often not successful for extracting interesting data. A PE decompiler or disassembler would be needed to extract this information. Right now there is no free software PE disassembler that is mature and easy to use.

Cygwin compliance engineering

Cygwin is a program which provides a real POSIX compliant system for Microsoft Windows. Cygwin is dual licensed under GPL. Red Hat also sells Cygwin under a proprietary license. A lot of GNU packages have been ported to Cygwin. Packages that need Cygwin to run have included a DLL (dynamic-link library) with POSIX compatibility code in it.

A Windows program can be easily detected using the 'file' command:

```
$ file a_program.exe
a_program.exe: PE32 executable for MS Windows (console) Intel 80386 32-bit
```

The contents of a file can be checked with the 'strings' program:

```
$ strings a_program.exe | grep cygwin
cygwin_internal
cygwin1.dll
_cygwin_crt0
__cygwin_crt0_common@8
_cygwin_premain3
_cygwin_premain2
_cygwin_premain1
_cygwin_premain0
__cygwin_crt0_bp
_cygwin_internal
_cygwin1_dll_iname
__head_cygwin1_dll
__imp__cygwin_internal
```

This is a clear indication that Cygwin is used.

Experiences

Experience from several years looking through several hundreds of source archives has learned there are a few easy targets to look for in GPL compliance. These targets can serve as a very simple litmus test for GPL compliance. One easy target is the toolchain. Often a binary-only toolchain is shipped in a GPL archive, without sources. For other tools, like the ones to create an actual file system (mksquashfs with or without LZMA compression, mkfs.jffs2, genromfs, mkcramfs) the sources are missing quite often too.

Another common violation is lack of bootloader sources (if a GPL licensed bootloader is used on the device) and add-on packages which were not part of the original SDK the vendor got from upstream.

A tricky source of violations, which is hard to explain to vendors, is when "extra software" is shipped in the GPL sources that is not present on the device. It often happens that a certain software stack for a particular board is used for developing various types of devices for various vendors. Traces of different devices, with different software, can show up in the GPL sources for a device, for example in the form of a file system with pre-compiled binaries, that was accidentally left in. While technically not interesting if you only want to tweak the software, this is a source for license violations. It is hard to explain to vendors, because in their eyes all the software that is on the device is in the GPL tarball, in a GPL compliant way.

Appendix A: GPL checklist

This is a small checklist for making sure a device is GPL compliant

1. Check the bootloader for GPL compliance. Use the list from the 'bootloader' section. If one of the bootloaders mentioned there is used, hunt down the sources.
2. Check if the device is GPL and LGPL compliant.
3. Check if the sources that are shipped are GPL and LGPL compliant (complete and not shipping more than necessary).
4. Check the documentation shipped with the device if it complies with GPLv2 section 1 and LGPLv2 section 1.

Appendix B: Reporting and fixing license violations

This guide presents some practical tips for solving common Free Software license compliance issues. It is not legal advice, and if in doubt, you should contact a qualified lawyer.

Reporting a violation

Be careful when reporting a violation. Accusations and suspicions voiced on public mailing lists create uncertainty and do little to solve violations. By checking your facts you can help experts resolve violations quickly.

Useful violation reports to companies about a potentially infringing product should contain:

- The name of the product affected
- The reason why a violation is believed to exist
- The name of the project code that may have been violated
- A statement regarding what licence this code is under
- A link to the project site

Useful violation reports to organisations like gpl-violations.org or the FTF should contain:

- The name of the project code that may have been violated
- A statement regarding what licence this code is under
- A link to the project site
- The name and website of the party who may be violating the code
- The reason why a violation is believed to exist

Additional tips:

- Please do not forward long email threads. They make it difficult to assess the situation.
- If you have clear evidence of a violation it is a good idea to tell the copyright holders. They can take legal action if necessary.

You can send violation reports to:

- gpl-violations.org: license-violation@gpl-violations.org
- FSFE's Freedom Task Force: ftf@fsfeurope.org

Handling a violation report

It is important to handle violation reports carefully. Free Software development focuses on community engagement and clear communication. That means it is important to respond to issues reported, even if your reply is initially brief. This helps prevent escalation.

Here are some useful steps:

- Confirm you have received any reports sent in and inform the reporter you are looking into the case
- If the report was made on a public forum try to move the discussion to a non-public space as soon as possible
- Isolate the precise problem. If you don't already have the information, ask the reporter for:
 - The name of the product affected or the exact code causing a problem
 - The reason why a violation is believed to exist
 - The name of the project code that may have been violated
 - A statement regarding what licence this code is under
 - A link to the project site
- Send updates to the reporter when they are available

Please bear in mind:

- Not every reporter understands licences fully and there may be mistakes in their submissions
- Compliance with the terms of the licences is not optional and lack of compliance can have serious consequences
- You can hire compliance engineers or purchase compliance services from third parties if necessary

You can get more information about best practice in this field by contacting:

- FSFE's Freedom Task Force: ftf@fsfeurope.org

You can obtain compliance engineering support by contacting:

- Loohuis Consulting: <http://www.loohuis-consulting.nl/GPL/>

Preventing a violation

The best way to fix violations is to prevent them occurring.

Useful tips:

- Read the licences you will use
- Check out the websites explaining these licences
- Get advice from experts

Useful tips for supply chain management:

- If third parties supply you with code, ensure you have licence compliance stipulated in your contracts
- Ask suppliers to bear the cost of resolving violations

For more information you can contact:

- gpl-violations.org: legal@lists.gpl-violations.org
- FSFE's Freedom Task Force: fff@fsfeurope.org

Copyright note

This appendix: copyright (c) 2008 Armijn Hemel, Shane Coughlan

This work is available under the [Creative Commons Attribution-No Derivative Works 3.0 Unported](https://creativecommons.org/licenses/by-nd/3.0/) licence.

Appendix C: Commercial compliance engineering

This documentation was made by Armijn Hemel at Loohuis Consulting, while doing research for gpl-violations.org.

Loohuis Consulting is specialized in tailor made hosting, development, training and consultancy.

Loohuis Consulting is one of the few companies in the world to offer GPL compliance engineering as a service. The increased use of Free Software requires an understanding of the licenses in use as well as best practice in deployment, deployment processes and compliance. Loohuis Consulting employees have practical experience in this field, especially with regards to embedded devices.

Loohuis Consulting is also one of the leading experts on Universal Plug and Play security. Our employees are pioneers in undertaking security audits on devices using Universal Plug and Play and have written award winning papers and given numerous presentations on the subject.

For more information please visit the Loohuis Consulting website:

<http://www.loohuis-consulting.nl/>