Mirroring CTAN

Michael Doob
Department of Mathematics, The University of Manitoba, Winnipeg, Manitoba, Canada

Abstract

The plummeting cost of disk space has made it perfectly reasonable to keep a copy of some or all of the CTAN archive on your local network or even on your own desktop computer. Keeping it up to date without creating unbearable traffic on your network is the job of mirroring software. We'll look at some software available that does this job (almost) painlessly.

The name of the game

Over the past decade CTAN (Comprehensive TEX archiving network) has become a standard tool for TEX users to obtain updates and even full installations of software and documentation. It is not only easy to access, but it also contains the latest version of (sometimes frequently) updated software. The only downside is that sometimes the internet can be congested and the hosting sites are quite heavily used. In addition, those not graced with high speed connections to the internet must show great patience when trying to download some of the larger packages. This situation can be ameliorated somewhat by using mirroring software.

The purpose of mirroring software is to keep a copy (mirror) of part or all of some authoritative site. This means that whenever a file is created, modified or deleted from that site, the same files is created, modified, or deleted in the mirror. This implies that the mirroring software must contact the site being mirrored on a regular basis to synchronize the file systems.

The current size of CTAN is just under 5 gigabytes, a size easily accommodated even on a personal desktop computer. In the next two sections we'll look at software using two different approaches to mirroring.

The wget program from the gnu collection

The program `wget` is a relatively small C program developed as part of the GNU project of the Free Software Foundation. The home site for the GNU project is http://www.gnu.org; the source code is available there. For most UNIX systems, it is simply a matter of going through the `./configure` → `make` → `make install` cycle. In addition, precompiled versions are available for the appropriate package manager for various Linux distributions (RedHat, Debian and Slackware) and `wget.exe` is on CTAN. Once the executable is available, using it for small sections of CTAN is almost trivial. Suppose you want to get the macros for TUGBoat. Then using the command `wget -m` ftp://ctan.tug.org/tex-archive/macros/latex/contrib/supported/tugboat will download all the files in the given directory on the (authoritative) site http://ctan.tug.org to your own computer. Of course you have to know the name of the directory you want (which can be pretty long) so it is natural to use `wget` in script or batch files. The `-m` option is for mirroring. This means that a record will be kept of the files downloaded, and calling `wget` again will result in only new or changed files being sent.

Once `wget` completes its download, you'll have a new directory http://ctan.tug.org which will have a subdirectory `tex-archive` and so forth so that the downloaded files will be in the same place in the directory tree as on the original site. Other pieces can be filled in as needed. In principal the whole site could be downloaded, but this would take days, even with a high-speed connection to the internet (see the following section for an alternative).

The `wget` program will automatically do recursive downloads of directories. The command `wget -mq` ftp://ctan.tug.org/tex-archive/macros/latex/contrib/supported would download all of the supported LATEX macros (this makes for a pretty big transfer!) The `-mq` option is for mirroring quietly so no informative messages come to the screen. Rerunning the command would download only the macros that changed or were added since the last download, so this could be done periodically (with a cron job in UNIX) and not incur long download times.

Michael Doob

It should be noted that if files are deleted on the site being mirrored, they are not deleted from your computer (so in a sense the mirror is half-silvered). There are also many other options that can be used with `wget`, and they are given in detail in the excellent manual available from the GNU site.

The `mirror` program from Imperial College, London

The `mirror` program is a perl script, so there is nothing to compile. It is available from http://sunsite.org.uk/packages/mirror or ftp://ftp.cs.columbia.edu/archives/perl/mirror. The main program is `mirror.pl`; it uses an auxiliary file `mirror.defaults` to configure itself. This is a text file pretty straightforward (and well documented) and allows, for example, for some files to be skipped. Here is a fragment to indicate the type of instructions that appear (in this example the CD images of the `texlive` distribution and the unneeded `00Contents` files are skipped):

```
package=ctan
   comment=Comprehensive TeX Archive Network (CTAN) at St. Marys College
   site=ctan.tug.org
   timeout=400
   exclude_patt+|\.texlive\*|00Contents|
   local_dir=/mirror/ftproot/tex-archive
   remote_dir=/tex-archive/
   max_days=0
   max_delete_files=20%
   max_delete_dirs=20%
```

When the `mirror` program is run, it downloads the current directory tree of the authoritative site and compares it with the mirror directory tree. It then downloads new or revised files from the authoritative site and deletes files on the mirror as necessary.

This program is robust enough to mirror the full CTAN collection. It might be preferable to copy the CTAN image from the available CD and then update from an authoritative site rather than download 5 gigabytes across the net. Once it is set up, a nightly run using `mirror` will normally update the mirror in less than half an hour.

Authoritative sites

There are three equivalent fully authoritative sites from with CTAN may be mirrored:

- ftp://ftp.dante.de (Hamburg, Germany)
- ftp://ftp.tex.ac.uk (Cambridge, UK)
- http://tug.ctan.org (Colchester, Vermont, USA)

In addition there are several dozen mirror sites listed in the file `/tex-archive/CTAN.sites` on CTAN itself (which themselves can be mirrored).

With only a small amount of effort, the areas of CTAN of interest can be mirrored on a personal computer; the latest software is then always available in a quick and easy fashion. It makes it easy to go forth and multiply!