Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)
March 20–22, Vienna, Austria *ISSN 1609-395X*
Kurt Hornik, Friedrich Leisch & Achim Zeileis (eds.)
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Quantian: A Scientific Computing Environment

Dirk Eddelbuettel
edd@debian.org

Abstract

This paper introduces Quantian, a scientific computing environment. Quantian is directly bootable from cdrom, self-configuring without user intervention, and comprises hundreds of applications, including a significant number of programs that are of immediate interest to quantitative researchers. Quantian is available at http://dirk.eddelbuettel.com/quantian.html.

1 Introduction

Quantian is a directly bootable and self-configuring Linux system on a single cdrom. Quantian is an extension of Knoppix (Knopper, 2003) from which it takes its base system of about 2.0 gigabytes of software, along with automatic hardware detection and configuration. Quantian adds software with a quantitative, numerical or scientific focus such as R, Octave, Ginac, GSL, Maxima, OpenDX, Pari, PSPP, QuantLib, XLisp-Stat and Yorick.

This paper is organized as follows. In the next section, we introduce the Debian distribution upon which Knopppix and, thus, Quantian are built, discuss Debian packages and its packaging system and provide an overview of the support for R and its related programs. We also describe the Knoppix system and provide a basic outline of Quantian. In the ensuing section, we describe the Quantian build process. Possible extensions are discussed in the following section before a short summary concludes the paper.

2 Background

Quantian builds on Knoppix, which is itself based on Debian. We will discuss these in turn, with a particular emphasis on support for R and related software.

A brief introduction to Debian

Debian is a Linux distribution. It was started around 1993 by the Free Software Foundation with a mandate to build a Linux distribution based on *Free Software*.[1] This principle is still at the core of Debian: all the software contained in a release is *free* according to the Debian Free Software Guidelines, or DFSG.[2]

Debian has grown from one to several hundred developers most of which are volunteers. Debian comprises a vast amount of software, distributed as so-called packages, which are self-contained binary archives. There are about six thousand source packages covering everything from well-known programs such as the Apache webserver, several Emacs and Vi editors, scripting languages such Perl or Python, compilers for at least a dozen different languages to less well known utilities; see MacKinnon (1999) for a review. From these sources, about eleven thousand binary packages are compiled. Each package is generally the responsibility of one developer, the package maintainer. Packages can enforce soft and hard constraints during build and install time – in other words, it can be ensured that while a package is built, another library providing a desired functionality ought to be present in order for its features to be reflected in the package. In such a case, installation of the package would also ensure that the library, if required rather than desirable, would be present at run-time. For many libraries, or even programs, multiple versions can be installed at the same time without conflicts. Packages can be installed, removed or upgraded using essentially a single point of control and interaction, the `apt-get` program which functions in a seamless and robust manner even across major Debian upgrades. It has been said that both the sophistication and reliability of this package management system are second to none compared to either commercial or Open Source variants of Linux or Unix.

Quality assurance also works at the package level. Flaws or bugs are reported on a per-package level, and are then in turn addressed by the package maintainer who will often release a revised version.

Debian is currently being prepared for eleven different hardware platforms six of which were included in the last release. These platforms cover everything from small processors such as the Arm (now used on PDA-size handhelds), common Unix architectures such as Alpha, Sparc or PowerPC to 64-bit processors such as Intel's ia64 as well as IBM s390 mainframe systems.

As Debian is freely available, the notion of 'market share' is hard to determine. However, informal surveys indicate a fair to high market (and mind) share among experienced users, yet a lower market share among novices. This lower usage among beginners is often attributed to what is said to be a more difficult initial installation compared to other Linux distributions. This very point is addressed squarely by the Knoppix (and thereby Quantian) bootable cdroms, which can turn essentially any cdrom-bootable PC into a fully configured workstation in just over a minute.

[1] The notions of *Free Software* and *Open Source* are defined by, respectively, the Free Software Foundation's GNU Project (`http://www.gnu.org`) and the Open Source Initiative (`http://www.opensource.org`). For the purposes of this paper, we will use the term Open Source though software contained in Quantian satisfies both criteria unless noted otherwise – it is both Free Software and Open Source.

[2] The Debian Free Software Guidelines are outlined at `http://www.debian.org/social_contract#guidelines`.

R under Debian

The R statistical language and environment (Ihaka and Gentleman, 1996) has been available for Debian since the 0.61 release in 1997. Debian provides extra features not present in other Linux distributions such as transparent support for the high-performance Atlas libraries (which can increase the speed of certain linear algebra operations by up to a factor of ten).

Debian also provides related packages such as the ESS mode for Emacs editors, and the XGobi data visualization program and its successor GGobi. Unfortunately, these two programs are released by AT&T under terms that are too restrictive for Debian to meet the DFSG and are therefore relegated to the non-free archive, which, strictly speaking, is not part of Debian but made available for the benefit and convenience of Debian users.

More recently, a few R packages from the Comprehensive R Archive Network (CRAN), the related Omegahat Project, as well as from contributing authors have been made available as Debian packages. Going forward, it is desirable to release additional packages which require either a more complex build process, external libraries such as database connection packages, or the XML support packages. Simple packages containing only R code, or a mixture of R and C code without requiring external libraries or header files, are easy to install via `install.packages()`. As well, Debian packages from the related BioConductor project for bioinformatics have been built, but are not yet widely distributed.

Knoppix

Knopper (2003) has built upon Debian by further integrating it into a so-called live cdrom that is directly bootable into a running system. One of his key engineering accomplishments was to combine the Linux loop device (used to mount cdrom media) with on-the-fly decompression. Thus, around 700 megabytes of compressed data on the cdrom correspond to around 2.0 gigabytes of uncompressed data. This allows a running Knoppix system to draw from the equivalent of almost three cdroms full of software even though only one is actually used. Knoppix has been available since 2000. At the time of this writing, version 3.2 is the most recent release.

Knoppix has to be seen to be fully appreciated as a rather sleek appearance is coupled with phenomenal hardware support and auto-detection. Virtually every type of computer, be it a server, desktop or laptop is detected with essentially all features – graphics, sound, networking (incl. wireless) and peripheral devices such as USB. With this automatic detection and setup, Knoppix goes from zero to a fully configured Linux workstation in around 1.5 minutes.

Knoppix has at least a dual focus of providing generally useful applications (such as OpenOffice or the KDE Office suite), games and multimedia players (such as Xmms) along with dozens of 'swiss-army knife' rescue, data recovery and networking tools appreciated by system administrators. It also comes with a variety of internet access tools allowing for network connections via wired or wireless ethernet as well as PPP or ISDN to either local or remote networks.

However, virtually no numerical or scientific software is included. For the scientific user, this is an obvious shortcoming, and where Quantian comes in as described in the next section.

Quantian

Quantian adds a numerical or quantitative bent to Knoppix. It starts with a normal Knoppix release which is then altered to make it more suitable for applied researchers. We will discuss the technical aspects of the remastering process in the next section, and focus in this section on a higher-level view.

Quantian, which is available at http://dirk.eddelbuettel.com/quantian.html, starts with the premise that Knoppix is already a very complete and successful product. Therefore, the intent is to remove none of the functionality providing automated hardware detection and configuration (of networking, graphics, sounds, etc.) at startup. Rather, we remove only a few programs which, while generally useful, are of more limited use in the context of data analysis or numerical work. One candidate for removal, primarily due to its sheer size, is OpenOffice, in particular given that two other spreadsheets (Gnumeric, Kspread), a work processor (Abiword) and several scientific publishing systems (Lyx, TeXmacs, XEmacs / AucTeX) are part of Quantian. We also remove a few other programs where the process used for the initial Quantian release is still somewhat eclectic. Among the removed programs are packages providing internationalization support, the Wine emulator, some supplemental Gimp packages, Abiword, ISDN and PPP networking, and some large games like Freeciv. The exact determination of what should remain, and what can be removed, should in the medium-term be driven by a mixture of user feedback and possibly some benevolent dictatorship applied by subcategory editors.

To this reduced set of packages, we add packages of interest to data workers and scientific or quantitative users. One set of packages is centered on the R language and environment and includes, beyond the R program and documentation packages, several add-on packages from the CRAN archive, the powerful ESS support for the family of Emacs editors, as well as the Ggobi visualization package. We also add Octave (an environment 'not unlike' a well-known commercial matrix laboratory) along with several of its add-on packages, the GSL (GNU Scientific Library) numerical libraries, the Maxima, Pari and Ginac computer-algebra systems, the QuantLib libraries for quantitative finance, the IBM OpenDX visualization toolkit, PSPP, XLisp-Stat and Yorick. We discuss possible extensions further below.

3 Mastering Quantian

Knoppix is built using the Debian distribution, and provides a Debian system when booted. For this reason, altering a Knoppix system requires some familiarity with the Debian package management system. General Linux administration knowledge is also helpful for the understanding of how to mount or write cdrom images in the typical ISO9660 format.

Knoppix customizations such as Quantian are becoming increasingly popular.[3] Essentially, the process of creating a Quantian cdrom involves the three steps outlined below.[4]

Copy setup from existing Knoppix cdrom: Insert Knoppix cdrom and mount it at, for example, `/cdrom`.

[3] A list can be found at the Knoppix support forum at http://www.knoppix.net/docs/index.php/KnoppixCustomizations.

[4] Practical matters of Knoppix remastering are discussed in the HOWTO guides http://www.stirnimann.com/mystuff/doc/knoppix.txt and http://tldp.org/LDP/LG/issue87/sunil.html.

Using the cloop module (which on Debian entails installing the cloop kernel module source package and building a binary kernel module package from it), the compressed Knoppix image can be mounted via `insmod cloop file=/cdrom/KNOPPIX/KNOPPIX`. This provides the decompression on-the-fly on top of the loop device used to mount ISO images. Next, assuming that a directory `uncompressed` has been created in the home directory, `mount -ro /dev/cloop ~/uncompressed` makes the approximately 2gb of software accessible below this directory in read-only form.

The content can now be copied to a new directory, say `~/source`, with `cp -Rupv ~/uncompressed/* ~/source/`.

As the content of the original Knoppix cdrom has been copied, we can remove the mounted drives: `umount /dev/cloop; rmmod cloop; umount /cdrom`.

Chroot and modify: We can now chroot into the expanded content which corresponds to making this directory the root directory of a new virtual session: `chroot ~/source`. This chroot-ed system requires a small amount of editing in `/etc/mtab`, `/etc/resolv.conf` and `/etc/apt/sources.list` so that networking is possible in order to obtain new Debian packages. A local NFS-mounted archive could also be used.

At this point, `dpkg` and `apt-get` can be used to tailor the 'running' system by removing and adding packages as discussed above. Further customizations and changes to the Knoppix scripts could also be made at this point.

Recompress and write to cdrom: Once the system is ready, a compressed file system can be made by piping the output of the standard `mkisofs` program through the script `create_compressed_fs` (which is included with Knoppix) and directing its output to a new file `release/KNOPPIX/KNOPPIX`.

The toplevel of the `release` directory has to be complemented with the other toplevel files obtained in step 1. The resulting filesystem structure can then be transformed into an iso file suitable for cd-writing using `mkisofs`.

Finally, the new iso image can be written with `cdrecord`.

4 Extensions

The schematic description in the previous section has shown how to build a Quantian cdrom. In this section, we describe possible extensions.

From the viewpoint of the *Distributed Statistical Computing 2003* conference and proceedings, the two most natural extensions are the inclusion of BioConductor as well as support for different database backends. Initial Debian packages have been produced from the BioConductor packages; these should install relatively easily into Quantian and thus make Quantian a valuable tool for bioinformatics. Database support is currently in a hold pattern until the R DBI initiative progresses further in terms of providing code for the agreed-upon DBI standard.

As far as standard Debian packages are concerned, Atlas is a prime candidate for inclusion. Atlas, whose acronym expands to Automatically Tuned Linear Algebra Software, is a system for the automatic generation and optimization of numerical software for processors with deep memory hierarchies and pipelined functional units. Atlas provides linear algebra libraries which tune extremely well for specific

CPU architectures. On medium to large matrices, speed increases by several orders of magnitude can be obtained just by linking to Atlas rather than standard libraries. However, this gain is obtained at run-time by matching the architecture of a given computer with a matching library package. Quantian, however, must load all possible versions in order to accommodate different hardware platforms, and then determine at boot time which one to activate. This should be relatively straightforward in terms of scripting support. In addition, the list of other packages from the math and science sections of the Debian distribution should be examined for other candidate packages for Quantian.

One of the most promising ideas for extension concerns computing clusters. A Quantian system could be made into a node for openMosix – this would allow for the rapid creation of ad-hoc computing clusters simply by booting a few machines on a LAN into Quantian cdroms, and providing a master host. This would have obvious appeal for anybody making use of 'embarrassingly parallel' methods such as Monte Carlo simulation or bootstrapping. Some conventions in terms of subnetwork addresses need some thought, but simply assuming the same subnet as the DHCP assigned address may be a start. Of course, this also requires replacing the Knoppix kernel with one that has the openMosix patch added onto it. Similarly, support for clustering from R, for example via the SNOW package, should be added given that Debian already has support for each of the three common parallel programming libraries systems MPI, LAM and PVM.

Lastly, the structure of Quantian may evolve along the lines of the modular Morphix variant of Knoppix. Where Knoppix and thus Quantian are monolithic in the sense that they provide a single large iso file, Morphix differs by offering different 'base', 'main', and 'mini' modules. Conceptually, Quantian could be split into one or more main modules for generally useful quantitative software, as well as domain-specific mini modules for specific scientific disciplines, or tasks.

5 Summary

We have outlined the Quantian scientific computing environment, a live system on a cdrom which can turn virtually any recent cdrom-bootable PC into a fully-configured scientific workstation in a matter of seconds. Quantian should prove useful for normal workstations, harddisk-less machines, teaching labs, demos, add-ons cdroms for books as well as for genuine experimentation, research and development. Quantian is available at http://dirk.eddelbuettel.com/quantian.html.

Acknowledgements

Comments by Lisa Powell, Tony Rossini and an anonymous referee are gratefully acknowledged, as are suggestions from both conference participants at DSC 2003 and from early Quantian users.

References

BioConductor. http://www.bioconductor.org, 2003.

Debian. http://www.debian.org, 2003.

Dirk Eddelbuettel. Quantian. http://dirk.eddelbuettel.com/quantian, 2003.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

Klaus Knopper. Knoppix. http://www.knopper.net/knoppix/index-en.html, 2003.

James G. MacKinnon. The Linux operating system: Debian GNU/Linux. *Journal of Applied Econometrics*, 14(4):443–52, 1999.

Morphix. http://morphix.sourceforge.net, 2003.

openMosix. http://openmosix.sourceforge.net, 2003.

R Project. http://www.r-project.org, 2003.