Quantian as an environment for distributed statistical computing

Dirk Eddelbuettel
dirk@eddelbuettel.com

April 14, 2005

Abstract

Quantian, initially introduced at DSC 2003 in Vienna and available at http://dirk.eddelbuettel.com/quantian, provides a comprehensive scientific computing environment in one bootable dvd. Using the advanced and automatic configuration system provided by Knoppix, Quantian is able to turn just about any standard laptop, desktop or server into a scientific workstation complete with approximately six gigabytes of software.

Of this readily useable scientific software, four gigabytes are scientific or quantitative in nature. The applications range from computer algebra systems to data visualization packages; they also include many domain-specific scientific application. Also provided are scripting environments (Python with dozens of add-ons packages, Ruby, Lua, Lush and of course Perl) as well as compilers, numerous libraries, debuggers and editors. Standard office tools and very extensive LaTeX tools provide excellent support for scientific publishing. Of course, the R environment has always been a always been a core part of Quantian, and several CRAN and Omegahat packages were included from the start. More recently, practically all packages from the CRAN and BioConductor repositories have been included making Quantian a particularly appealing choice for those wanting to with R or BioConductor.

One of the key extensions to Quantian has been the addition of an openMosix-enabled kernel as well as userspace tools for openMosix. The openMosix kernel, together with the ‘terminalserver’ capabilities in Quantian, can turn any standard computer lab into a supercomputer. Simply booting one ‘head node’ off the dvd and enabling the network booting allows to boot additional nodes (which could be diskless) off the head node. Network booting ensure that all nodes share the exact same kernel version and configuration and thus pass the core requirement for joining an openMosix cluster. Such a single-system image (SSI) cluster permits the user to employ potentially dozens or even hundreds of standard PCs as if it were one large multi-processor computer which has obvious appeal for simulation-based methods of scientific inquiry.

Moreover, Quantian also includes full support for ‘classic’ distributed / parallel computing using the LAM/MPI and PVM toolkits. These are available directly (using explicit C/C++ or Fortran programming), or at a more abstract level in R using the SNOW package by Tierney et al. The SNOW package employs random-number generators suitable for parallel streams (SPRNG is the current default, the rlecuyer RNG may be added) making this package ideal for high-level simulation-based methods such as bootstrapping or markov chain monte carlo.

A key advantage of Quantian in the domain of distributed statistical computing is that the openMosix kernel extensions can also control the MPI or PVM process migration. After starting a large number of (either MPI or PVM) processes on one node, openMosix will transfer processes to less-utilised nodes in the cluster which ensures an efficient use of resources without the user having to explicitly control the migration, or even define, control or update the set of available nodes.

We plan to illustrate several of these key features of Quantian with short examples.