Astronomical Data Analysis Software and Systems XVI **P3-11**
ASP Conference Series, Vol. XXX, 2006
R. Shaw, F. Hill and D. Bell, eds.

Astronomy, Linux & The Problem of Large File Support

Ross Collins, Johann Bryant, Nicholas Cross, Nigel Hambly, Mark Holliman, Mike Read, Eckhard Sutorius, Peredur Williams

Wide Field Astronomy Unit, Institute for Astronomy, University of Edinburgh, UK

Abstract. A review of the current status of support for large FITS files (those greater than 2 GB) and 64-bit PC architecture in Linux software applications commonly used for astronomical data processing. We present solutions for the support of large FITS files on 32-bit Linux systems, together with a report on the status of 64-bit support in the Starlink and Python libraries used to process FITS files on Linux PCs. Our conclusion is that 64-bit astronomical software for Linux is now sufficiently mature that data centres can now confidently migrate from 32-bit PCs to 64-bit PCs to overcome various technical limitations of 32-bit PCs.

1. Introduction

Large astronomical surveys such as the UKIRT Infrared Deep Sky Survey (UKIDSS, see Lawrence et al. 2006) are now not only generating terabytes worth of data, but their pipelines are also producing individual files that exceed the 2 GB limit of 32-bit Linux PCs. Mosaic FITS files for the WFCAM Science Archive (WSA, see Hambly et al. 2007, also Collins et al. 2006), for example, will potentially reach 15 GB in size. Although the C library for 32-bit Linux has, for a long time provided support for large files, only now are astronomy software applications catching up, and inevitably the problem of a 4 GB memory limit for 32-bit architectures also arises. The logical move towards a 64-bit architecture is hampered by currently patchy support in Linux for truly 64-bit applications. Thus this paper is split into two sections. In section 2. we discuss solutions for ensuring that your 32-bit Linux PC has large-file support in all FITS I/O applications. Section 3. reviews the current status of 64-bit support for Linux PCs, with an emphasis on the software used by the WSA, particular the Starlink software suite, and Python programming language libraries.

2. Large-file support

To ensure that your 32-bit Linux PC system can process FITS files larger than 2 GB in size:

- Make sure that C/C++ applications compiled with the *gcc* compiler define the preprocessor option `_GNU_SOURCE` or `_LARGEFILE_SOURCE`. This option may be specified at compile time in the Makefile, or specifically in the source header.

- Build the latest CFITSIO (currently version 3.02) and compile all application software against this version of the library. Since version 3.0 (released February 2006), the library now compiles with large-file support by default, whereas earlier versions require you to specify a preprocessor option upon compilation.
- Install the Keoe release of the Starlink software suite (released September 2006). This is the first version to officially support large-files. However, one still cannot view FITS images, in say GAIA (Draper 2007), that are larger than the currently available physical memory. For 32-bit PCs, physical memory is limited to 4 GB.
- When editing the headers of large FITS files using PyFITS one cannot create a new header block if the size of the file is greater than the currently available physical memory. Therefore, it helps to add some additional 'padding' to FITS headers to accommodate future needs.

In summary, large-file support for FITS files on 32-bit Linux PCs exists, but owing to the memory limitations mentioned above, the following functions cannot be performed:

- Viewing FITS files with size greater than the currently available physical memory (e.g. with GAIA or DS9).
- Adding new FITS header keywords that require the creation of a new header block on a file with size greater than the currently available physical memory (e.g. with PyFITS).
- Adding HDU extensions to a FITS file with size greater than the currently available physical memory (e.g. with PyFITS).

3. 64-bit support

A transition towards the 64-bit architecture removes the fundamental 4 GB limit upon the PC's physical memory. However, software for 64-bit Linux has been slow to mature, most notably resulting in the most recent Debian distribution (*Sarge*) not providing support for 64-bit PCs. Recently this situation has shown notable improvements, with the Python programming language now fully supporting 64-bit architectures (as of version 2.5, released September 2006) and Debian announcing that their next release (*Etch*) will be distributed with a fully supported 64-bit version.

Astronomy software for the most part also now provides 64-bit support, and the recent (September 2006) Keoe release of the Starlink software suite provides 64-bit binary packages. The Python PyFITS module to date has relied upon the numarray library, which is inherently 32-bit. The latest development version (1.1, which had a beta 3 release in September 2006) is compatible with NumPy (numarray's successor; with a 1.0 version release in October 2006), which by replacing numarray provides PyFITS with full 64-bit support.

Using an AMD Opteron 64-bit PC with 4 GB of physical RAM, running the 64-bit version of the OpenSUSE 10.1 Linux distribution, we performed the same set of tasks (as tested in the previous section) upon our suite of Starlink and Python based software. We were able to view FITS files of any size up to the currently available amount of physical memory (with e.g. GAIA, DS9). Furthermore, it was possible to update the headers, add any number of new FITS header keywords, and create as many new extensions as we like for FITS files of *any* size. There are no memory limitations. However, with PyFITS, if the file is greater in size than the available physical memory, append-type alterations fail (where the entire file must be read into memory, say due to the creation of a new header block), so you must write the changes to a new file.

4. Conclusion

Many astronomical data analysis centres rely upon 32-bit PCs running Linux to process FITS files. As new telescopes with larger detector arrays begin to produce individual images greater than 2 GB in size, it is clear that 32-bit PCs will soon no longer be adequate for the task of data analysis. However, as our tests for the WSA have shown, 64-bit versions of both Linux and astronomical software are now beginning to become sufficiently mature to allow data centres to perform the migration from 32-bit PCs to 64-bit PCs.

References

Collins, R.S, et al., 2006, in ASP Conf. Ser., Vol. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique (San Francisco: ASP)

Draper, P., 2007, this volume, [D1.1]

Hambly, N.C., et al., 2007, in preparation

Lawrence, A., et al., 2006, MNRAS, submitted (astro-ph/0604426)