



Astronomy, Linux and The Problem of Large File Support



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

Wide Field Astronomy Unit (WFAU), Institute for Astronomy, University of Edinburgh

Introduction

Large astronomical surveys such as the UKIRT Infrared Deep Sky Survey (UKIDSS) 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, 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. Here we review the current status of large file and 64-bit support in Linux software applications commonly used for astronomical data processing.

The Problem

Your sky survey data pipeline is beginning to produce FITS files greater than 2 GB in size, such as this mosaic image for the UKIDSS Ultra Deep Survey



The Solutions

Large File Support for 32-bit PCs

64-bit PC Support

Linux



All major Linux distributions have x86-64 support, with the significant exception of Debian, which will not provide x86-64 support until the release of version 4.0 (*etch*) in Dec 06 (?!).

C & C++ Compiler

GCC requires software to define the preprocessor option `_GNU_SOURCE` or `_LARGEFILE_SOURCE`, which may be specified at compile time in the Makefile.

CFITSIO Library

Versions > 3.0 (released Feb 06) now compile with large file support by default, whereas earlier versions require you to specify preprocessor option on compilation.

Versions > 3.0 (released Feb 06) supply separate functions specifically to provide support for 64-bit integers.

Starlink

The Keoe release (Sep 06) provides large file support. However, cannot view FITS images in say, GAIA, that are larger than the available physical memory, which is limited to 4 GB!

The Keoe release (Sep 06) comes with a working 64-bit version. Viewing of FITS images in GAIA is still limited by the available physical memory.

Python & PyFITS



PyFITS supports large files. However, edits to FITS files larger than the available physical memory can only be saved if the file size remains the same size (i.e. a new header block, or HDU does not need to be created). Memory mapping does not work in this "append" mode, and is also bound by 32-bit limit.

Python 2.5 (released Sep 06) provides full 64-bit support, using 64-bit integers for all indices.

PyFITS 1.0 does not have 64-bit support as it relies upon the 32-bit numarray library.

However, PyFITS 1.1 (*beta 3* released Sep 06), is compatible with NumPy (*release candidate 1* released Sep 06), which supports 64-bit.

For files larger than the available physical memory, "append"s fail, so the edits must be saved into a new file.

This investigation was carried out as part of a design review for the WFCAM & VISTA Science Archives (WSA & VSA). For more information on the WSA & VSA please see:

Python & C++ Software Documentation: <http://www.roe.ac.uk/~rsc/wsa/>
WFCAM Science Archive Design Documents: <http://www.roe.ac.uk/~nch/wfcam/>
Web interface to the WFCAM science archive: <http://surveys.roe.ac.uk/wsa/>

