HENSA Unix – Serving the UK

Maggie Bowman, Tim Hopkins and Neil G. Smith
Computing Laboratory
University of Kent at Canterbury
Canterbury
Kent CT2 7NF

February 1995

Abstract

Following a short evolutionary history of the HENSA Unix service, we look in some detail at three of the newer facilities being offered, the World Wide Web proxy cache, indexing of the archive, and the bibliographic service. We describe briefly the current contents of the software repository which has, by user demand, been extended to mirror more sites. Finally we present some usage statistics and comment on the way in which access to HENSA Unix has changed recently.

Introduction and Background

HENSA Unix is a JISC funded, national service run from the Computing Laboratory at the University of Kent. It is part of the Higher Educational National Software Archive which includes HENSA Micros at the University of Lancaster. Both HENSA services are available free to all members of the UK Higher Education community.

The HENSA Unix service grew out of an SERC funded pilot scheme to provide a UK netlib [1] server at Kent. The original netlib service at AT&T allowed research workers to obtain high quality scientific software via electronic mail and the provision of a UK server was aimed at reducing the need for expensive accesses to the US server. This UK pilot server was expanded and diversified. Users could execute programs on the netlib machine using their own data and have the results returned to their mailbox. Material from other sources was also made available. The means of access was extended to allow ni-ftp (Blue Book) and later anonymous ftp. New system software, written at Kent, improved the servicing of user requests and the returning of material via electronic mail.

Initially the state of the art, as far as repository mirroring was concerned, was either to exchange tapes or to attempt to keep the archives in step by requesting changed material by electronic mail. With the arrival of an Internet link at Kent a shadow copy of the uunet archive at ftp.uu.net was automatically kept up-to-date, again using locally written software. The emphasis was now very much on Unix based material though uunet also provided a wealth of general interest information. The availability of the uunet archive changed the perception of the service and attracted users from a wide variety of disciplines and interests.

Shortly after this expansion the ISSC identified the need for a national service to provide freely available software and associated information to the UK Higher Education community and HENSA was born. The arrival of a fast JIPS link at Kent enabled an interactive service to be introduced. Following demands from users more archive sites have been shadowed and additional means of accessing the software offered.

The services available via HENSA Unix include

- an archive of Unix related software including numerical and statistical packages, system software and development tools, newsgroups and general information mirrored from some 14 sites. Mirrored material is updated daily and the contents of the archive are all freely available. HENSA Unix is also host to a parallel computing, and a dyslexia information and software archive both of which are maintained by staff at the University of Kent,

- a World Wide Web proxy cache which allows users to obtain documents without the need to access overloaded servers across slow and busy international links,

- a bibliographic service which enables users to locate, preview and retrieve copies of technical reports and other documents maintained locally and at other UK and European sites. The database now contains 1239 documents from 16 sites,

- an archie server which allows users to perform searches for publically available material from sites across the Internet,

- a Special Interest Group Archive which offers SIGs such as AGOCG, the Computational Fluid Dynamics Group at RAL and the Graphics

Group at Manchester a central distribution site for their material. The contents of each SIG area of HENSA Unix is controlled by the individual groups.

Currently HENSA Unix supports access to the software archive via WWW, anonymous ftp, telnet (allowing interactive browsing of the contents of the archive and elementary searches on filenames), gopher and electronic mail.

The original netlib server ran on a Sun4 with 327 Mbytes of disk but was expanded to allow the service to be extended to include the uunet archive. The first move was to a Sparc2 with 5 Gbytes of disk and then to a SPARCstation 10 with 14 Gbytes of disk. At this stage the Sparc2 was used to run the HENSA Unix archie service. With the popularity of the Web cache and the resultant huge growth in the number of requests made to the server, the increase in data due to further sites being mirrored, and the need for more processing power to enable indexes to be built, a further upgrade is about to take place. The SPARCstation 10 is in the process of being replaced by a Silicon Graphics Challenge S with 256 Mbytes of main memory and 27 Gbytes of disk; this machine will take over the software mirroring role and the Web cache will have sole use of the SPARCstation 10.

In the following sections we look in detail at some of the newer services offered by HENSA Unix and give a short description of the content of the software repository. Finally we provide some usage statistics and comment briefly on the changing trends of access to the HENSA Unix facility.

The World Wide Web Proxy Cache

The philosophy behind the use of a WWW proxy cache and a discussion of the relative merits of caching and mirroring are given in Smith [3].[1] Browsers like Netscape may be configured to use the cache and this means that rather than going directly to the primary source for a document, the client requests it from the cache. If the document has been requested recently it may already be in the cache and can thus be delivered to the user without the need to use slow and overloaded international links. Should the document

[1] This is available from unix.hensa.ac.uk either by anonymous ftp as compressed PostScript in the file `/pub/misc/ukc.reports/comp.sci/reports/11-94.ps.Z` or via WWW using the URL `http://www.hensa.ac.uk/www94`

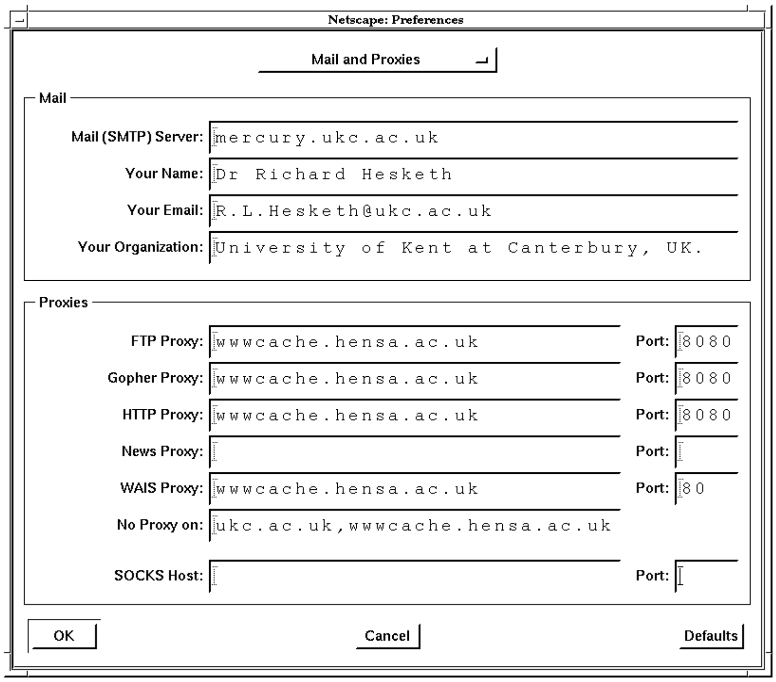

Figure 1: Configuring the Netscape Browser to use the HENSA Unix WWW proxy cache

not be available the caching software retrieves it, delivers it to the user and, usually, keeps a copy within the cache. Figure 1 gives details of how to configure the Netscape browser to use the HENSA Unix proxy cache.

The length of time that a document resides within the cache is determined by a number of factors. If a document includes an expiry date, that date is always respected. If there is no expiry date then heuristic rules are used to estimate the lifetime of a document. This is taken to be 10% of the time since the document was last modified up to a maximum of 12 hours. This means that a document that was modified 10 hours ago will only be kept for one hour before it becomes stale in a cache. Once a document has become stale a check is made against the primary source to see whether the file has changed and a fresh copy is fetched if it has.

The size of the HENSA proxy cache is currently restricted to 1.8 Gbytes due to lack of disk space on the server. Even so the hit rate, i.e., the percentage of the requests that result in documents being recovered from the cache, is around the 42% mark. Simulations based on the request logs indicate that the optimum cache size is around 17 Gbytes when the hit rate may be expected to rise to around 56%. Based on the cache use for February 1995 the 42% hit rate on around 3 million requests represented approximately 12 Gbytes of traffic. With an extended cache in place this could be expected to have been around 16 Gbytes. Considering the relatively small fraction of Higher Education institutions currently accessing the cache this could represent both a considerable saving in terms of international bandwidth and a substantial decrease in the time users have to wait to retrieve documents.

Indexing the Archive

The HENSA Unix Archive has simply grown and grown. The breadth of information available is overwhelming and finding the particular file or directory a user requires is now a real problem. The archive has always offered searches based on simple filename matches but this is often not enough and more sophisticated methods have to be provided.

One method of indexing the contents of the archive is to index the text files that can be found associated with a particular package. In general these files have fairly standard names (something involving README or INSTALL). A keyword search over this database may provide useful pointers. For example, a simple filename search for the word debugger would not find either ups or gdb, but this new indexing method should find both of them and more.

These free text style searches can be extended to other parts of the archive. For example, we have complete sets of USENET frequently asked questions, Internet RFCs, Project Gutenberg and OBI electronic texts. Making these files searchable does not necessarily make it more practical to find a file on the archive, but it is a valuable service that we will be providing.

There are limits to the amount of automatic indexing that can be performed. Sometimes it is necessary to have some intelligent human input. We hope that in the near future we shall be providing a thematic index to the whole of the archive. In the past we have always mirrored remote sites exactly and not attempted to arrange things thematically. We shall continue to provide these simple mirrors for the people who know exactly what they are want and where it is, but in addition provide a comprehensive catalogue for those who need to browse around a bit first.

```
@techreport{ukc11-94*,
        author         = "Neil G. Smith",
        title          = "What can Archives offer the World Wide Web?",
        institution    = "Computing Laboratory",
        year           = "1994",
        number         = "11-94*",
        address        = "University of Kent, Canterbury, UK",
        month          = jul,
        ftpaddress     = "unix.hensa.ac.uk",
        filename       = "pub/misc/ukc.reports/comp.sci/reports/11-94.ps.Z",
        abstractfilename= "pub/misc/ukc.reports/comp.sci/abstracts/11-94",
}
```

Figure 2: A Sample Extended Bibtex Entry

Bibliographic Service

This service offers research workers the ability to search a database of technical reports, theses and conference proceedings from a number of UK and European institutions using either gopher or WWW. From the list of reports displayed following a successful search, it is possible to request particular abstracts and reports. If WWW is used, and a suitable viewer is available, many of the reports themselves may be previewed since most are provided as PostScript files.

None of the reports retrievable are stored on the HENSA Unix machine apart from the University of Kent's own documents. Rather each participating institution is asked to provide an index file containing a Bibtex entry (see [2] for details of Bibtex) for each document the site wishes to make available. A sample entry is given in Figure 2. Abstracts may be included as text within the Bibtex entry or the name of the remote site file containing an ASCII version of the abstract may be provided. These index files are mirrored nightly and any changes are visible following this update. When a paper is requested an ftp transfer is initiated between the user and the remote site.

Approximately sixty of the technical reports and corresponding abstracts in the database are provided by the Computing Laboratory at the University of Kent. During the past four months there have been 500 requests for these reports and 1279 requests for the associated abstracts. Further information on how to include a document collection from a new institution into this service may be found in the files

```
/pub/info/tech.reports.info
```

and

```
/pub/info/tech.reports.online
```

available from unix.hensa.ac.uk or by contacting any of the authors. The aim is to make available a wide range of subject matter without concentrating solely on computing and numerical topics.

Archive Contents

The complete software repository consists of a number of mirrored international archives along with two that are maintained locally. All mirrored sites are updated daily and it is rare that material is more than a day out of phase with the master copy. In total the archive contains some 300,000 files and occupies about 12 Gbytes.

By far the largest of the individual mirrored sites is the uunet archive which contains a large general-purpose collection of sources and documentation for Unix and Unix-like systems together with non-Unix material of general interest. It is a particularly good source for Unix systems material including the GNU free software foundation sources, X11R6 and X applications, USENET's comp.sources and news archive, and a number of mail systems. However, these examples only represent a very small fraction of the wealth of software and information available. Users who are not familiar with the uunet archive are encouraged to browse around the system in order to appreciate the wealth of stored material.

The archive's pedigree is reflected in the large scientific software presence with no less than five subarchives covering this area. Netlib and Statlib cover the high-quality, general purpose numerical and statistical software areas along with datasets and documentation. The other three collections contain contributed material for popular commercial packages; matlab, a scientific computation and visualisation system, and the computer algebra systems, maple and mathematica.

Another strong area of the archive is the freely available Unix-like operating systems for PC's; linux and FreeBSD. Along with the basic operating system a wide range of utilities and applications, including emacs, X11, compilers, and networking and mail software, are also available.

The locally maintained parallel computing archive is funded by the JISC through the London and South East High Performance Computing Centre (SEL-HPC). It consists of around 200 Mbytes of freely distributable software, documents, papers and other material related to Parallel and High Performance Computing.

The creation of the dyslexia archive is being funded by the HEFCE under its Special Initiative for Widening Participation. The archive will cover all aspects of dyslexia including software packages and the interface is being carefully designed to make it accessible to dyslexic students.

It is intended that the archive will continue to grow to satisfy user demand for making high quality software available locally. The new server, with 27 Gbytes, has twice the disk space available on the current machine and there is provision to expand to 72 Gbytes.

HENSA Unix Usage Statistics

Since HENSA Unix went live in June 1992 there has been a steady increase in the number of transactions the system handles and the amount of data being transferred. Figure 3 shows a history of the service over the past two years. The introduction of WWW access to HENSA Unix and the provision of the Web cache make a clear impact in March 1994. The graph of the WWW cache, Figure 4, shows both the number of accesses and the number of cache hits for the last 25 weeks. Again this figure reflects the growth of use of WWW. The shift in the means of accessing HENSA Unix since the arrival of WWW is dramatic. Tables 5 shows the number and relative percentage of requests to HENSA Unix for each of the access methods available. Similar data for the volume of data generated is given in Table 6. These clearly show the growing popularity of the proxy cache and the use of WWW as a means of accessing the archive.

The amount of data transferred by electronic mail and gopher has remained almost constant between July 1994 and February 1995 but, as a percentage of the total data, these methods are becoming almost insignificant. On the other hand, the volume of traffic generated by anonymous ftp has steadily increased although its percentage share has decreased. All, however, have been completely outstripped by the growth of WWW which has now taken over as the primary means of accessing HENSA Unix.

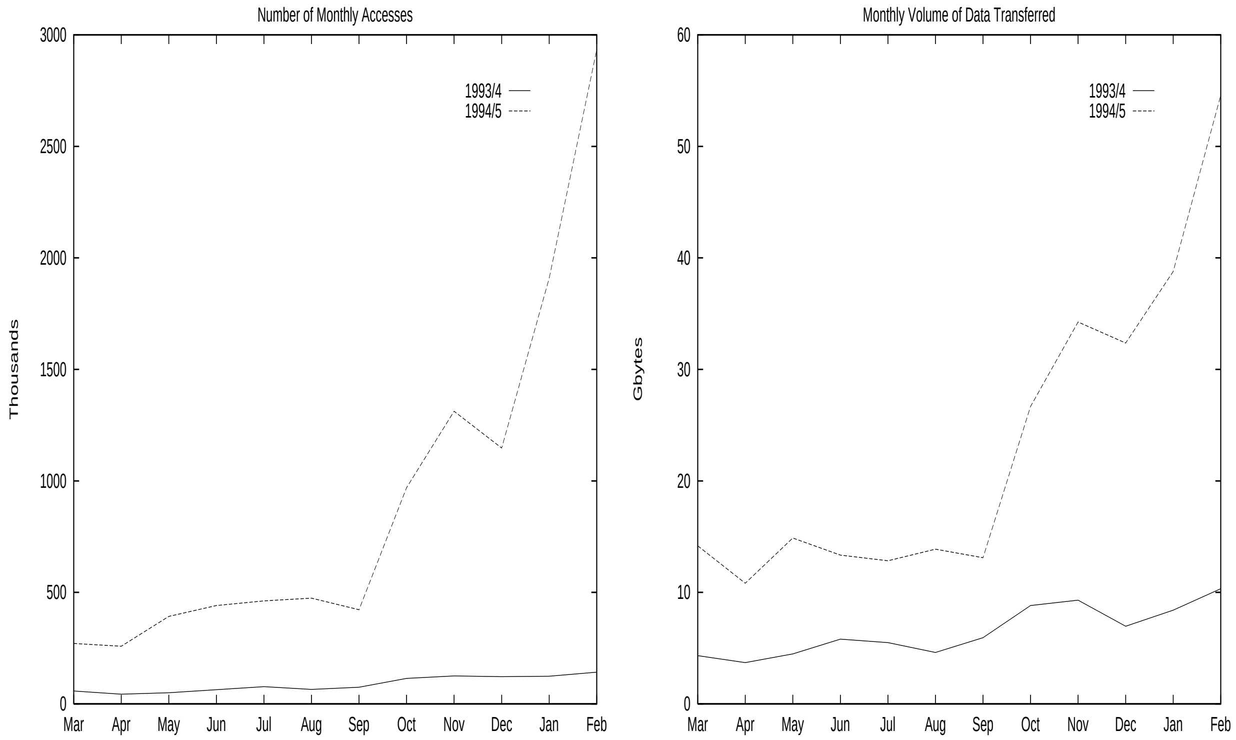

Figure 3: Transactions and data transmission for HENSA Unix

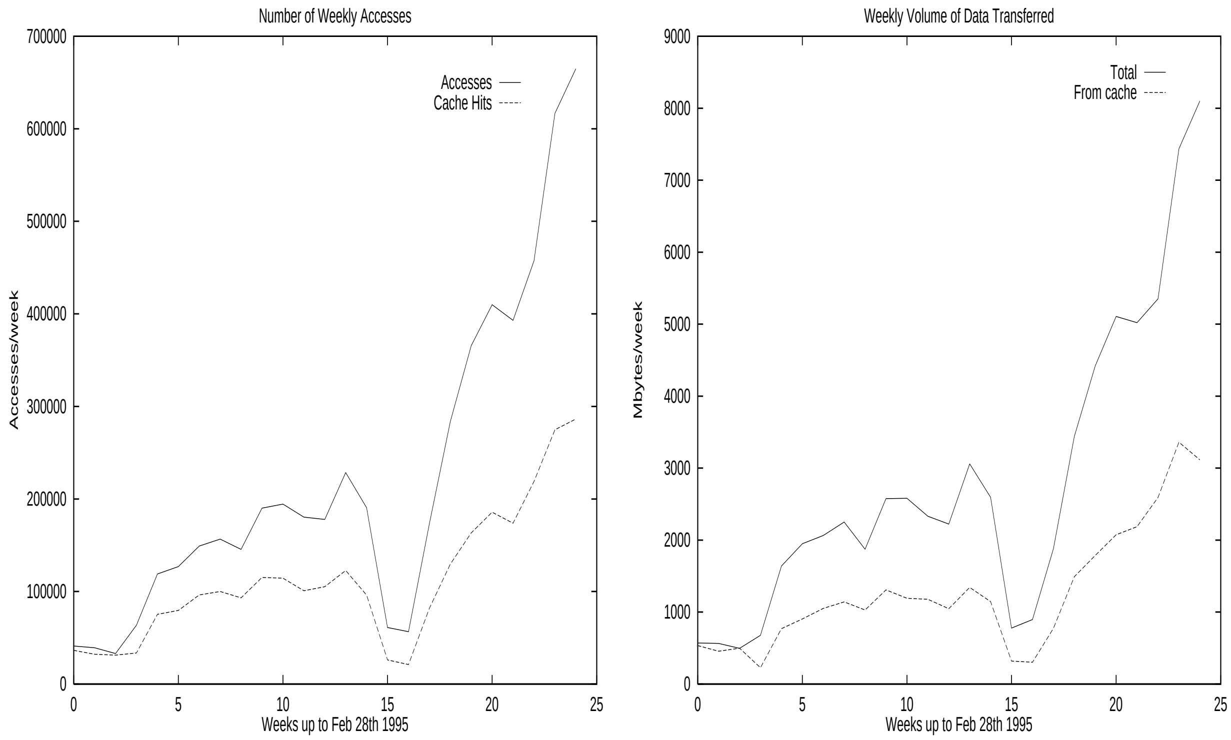

Figure 4: Transactions and data transmission for WWW proxy cache

	% Requests		Thousands	
	July 1994	Feb 1995	July 1994	Feb 1995
Electronic mail	0.38	0.09	1	1
Interactive Requests	0.32	0.06	1	2
Gopher	5.27	1.23	24	36
Anonymous ftp	39.56	10.17	183	298
World Wide Web	54.47	88.45	251	2594

Figure 5: Archive Requests by Access Method

	% Volume		Mbytes	
	July 1994	Feb 1995	July 1994	Feb 1995
Electronic mail	1.32	0.51	170	280
Interactive Requests	0.18	0.04	24	24
Gopher	4.23	1.12	54	60
Anonymous ftp	71.04	40.01	9117	21841
World Wide Web	23.23	58.32	2982	31833

Figure 6: Volume of Data Transferred by Access Method

References

[1] Jack Dongarra and Eric Grosse. Distribution of mathematical software via electronic mail. *Commun. ACM*, 30(5):403–407, May 1987.

[2] Leslie Lamport. *LaTeX User's Guide & Reference Manual.* Addison-Wesley, Reading, Mass, 1986.

[3] Neil G. Smith. What can archives offer the world wide web? Technical Report 11-94, Computing Laboratory, University of Kent, Canterbury, UK, July 1994.