VISTA DATA FLOW SYSTEM (VDFS)

---

for VISTA & WFCAM data

WSA Hardware/OS/DBMS design Document

author
E. Sutorius (WFAU Edinburgh)
Database Scientist
number
VDF-WFA-WSA-006
issue
Issue 1.0
date
2 Apr 2003
co-authors
N. Hambly, A. Lawrence, R. Mann, H. MacGillivray

Contents

1 SCOPE

This Hardware/OS/DBMS Design Document (HDD) for the WFCAM Science Archive (WSA) gives an overview of the hardware for data storage as well as for the data servers used for the WSA at the archive centre (WFAU at the IfA, Edinburgh). The reason for considering hardware, OS, and DBMS together is that the three are intimately linked; optimal DBMS operation is not always possible for a given hardware and OS configuration. Since copying the data from the data processing centre (CASU at the IoA, Cambridge) to WFAU plays a significant role, networking is also considered.

This HDD is intended to be a reference to software engineers and scientists working on the WSA project. A primary goal of the document is to specify the first phase WSA hardware in enough detail to enable orders to be placed with vendors to acquire the necessary equipment. This document will also discuss how we plan to move from V1.0 through the first year of data ingest to the V2.0 system.

2 INTRODUCTION

In our developments for the WSA so far, we have had to deal with some tensions. The need for timely development work (eg. hands–on experience) has to be balanced against the general best practice of delaying hardware purchases as long as possible (eg. Moore's law). Further, the timescale for significant developments in computer hardware technology implies that over the WSA lifetime the archive hardware solution is likely to change and a migration from one system to another will almost certainly be needed (this is even more so for the overall VDFS). Hardware configuration has many complicated variables, and it is only via experimentation that most questions can be answered.

Fortunately, hardware manufacturers (eg. Sun MicroSystems, IBM, Compusys) are open to donation of discounts, money and/or hardware for what they see as big projects in R&D. Furthermore, many hardware suppliers are open to loan of high–specification kit, meaning that experimentation is possible with no outlay on hardware. The commercial hardware/software big players (ie. MicroSoft/SQLServer, IBM/DB2 and Oracle) are open to supplying expertise and advice to large database developers and, more importantly, to supplying licences that are heavily discounted or even free–of–charge. Finally, many of the hardware issues overlap with similar ones in other IT projects in the UK (eg. initiatives at the Edinburgh National e–Science Centre, within AstroGrid, etc.) and farther afield in Europe (eg. Astronomical Wide–field Imaging System for Europe, ASTRO–WISE [1]; and the ESO Next Generation Archive Systems Technologies, NGAST [2]).

In this document the hardware components that will be used to build the WSA are described. In Section 4 we describe the required storage hardware for the pixel data archive and the catalogue data archive, and also describe the servers needed: the load server to efficently upload all the data and the public servers, which will handle the access to the WSA. The backup system is described in Section 6. Finally, Section 7 discusses infrastructure (local area network, equipment accommodation etc).

Applicable documents are listed in Section 10.

3 REQUIREMENTS

The clear requirement to arise out of the SRAD (AD01) is for a phased approach so that a WSA with 'standard' functionality is available at first light, enabling immediate science exploitation, and subsequently a fully functioning archive system is made available one year after survey operations begin in earnest. Furthermore, there is a clear split in the requirements for volume and access speed for catalogue and processed pixel data (hereafter, 'pixel data' will mean processed pixels; note that there is no requirement on the WSA to store the raw pixel data). The WSA usages require user interaction

with *catalogue data* (ie. complex queries returning results in as close to real time as is feasible) for data mining and data exploration while high volume *pixel data* usages are less time–critical (users would be prepared to wait for the results of operations on large pixel volumes since these would be executed relatively rarely).

The fundamental requirement on the WSA system hardware concerns the volume of data that will flow into the archive, and the rate at which that data flow occurs. In Appendix 8.1 we give an anlysis of data volumes and rates based around current knowledge and reasonable assumptions.

The outline hardware requirements are therefore as follows:

- V1.0 system by the end of Q4 2003 (to be ready for WFCAM first light – currently Q1 2004);
- V2.0 system one year after survey operations begin. Surveys will begin at the end of Q1 2004, hence V2.0 must be available at the end of Q1 2005 (note: acquisition of V2.0 hardware takes place in Qs 3 & 4 2004);
- on average, 100 Gbytes of data will be transfered from CASU for every night of WFCAM observations. If we further assume that the peak data rate could be as much as twice this figure, and require that the network transfer from CASU must take place overnight to avoid heavy network use during normal working hours – say 10 hours – then the required bandwidth is 100 Gbytes in 5 hours, or 6 Mbyte/s;
- WFCAM pixel volume: 20 Tbyte/year; speed of access is not a critical issue for large amounts of pixel data (eg. any large 'batch' pixel usage will not be time critical);
- Other pixel data requiring storage (see the SRAD) will be dominated by the SDSS pixels: DR1/DR2/final releases will be 15%/50%/100% complete where the total SDSS pixel volume is $\sim$ 10 Tbytes. V1.0 will include DR1 and V2.0 will include DR2; SDSS pixel volumes are 1.5 and 5 Tbytes respectively for DR1 and DR2.
- WFCAM object catalogues/ancilliary data: 2 Tbytes/year; 'real–time' access is required (ie. allow users to interact with and explore the data and we suggest $\sim$ 100s response time is therefore a reasonable goal);
- Other catalogues requiring storage: again, see the SRAD; they will be dominated by the SSS & SDSS each of which is of order 1 Tbyte.
- scalability (from V1.0 to V2.0 and beyond to VISTA): clearly, every year of operation will accumulate another $\sim$ 22 Tbytes of WFCAM data, but the scalability requirement is not simply one of increased storage capacity. Catalogue curation will become more time consuming as more data accumulate, so the hardware/OS/DBMS programme must take account of this;
- security: the SRAD requires data to be easily and quickly recoverable in the event of accidental loss.

Hence the split is V1.0/V2.0 and pixels/catalogues. After one year of operation, the V1.0 catalogue volume is 5 Tbytes; the V1.0 pixel volume is 22 Tbytes. Data accumulation is then 22 Tbytes/year (pixels) and 2 Tbytes/year (catalogues), so after one year of operation of the V2.0 archive the pixel volume will be larger by 25 Tbytes (including SDSS DR2) and the catalogue data volume will be larger by 2.5 Tbytes (including SDSS DR2).

The phased approach and volume/speed split is reinforced in the light of the hardware considerations stated in the previous Section: we will implement two distinct hardware solutions to satisfy the user requirements. If one subsequently becomes a viable solution for both, but at the same time at reduced cost, then migration from one solution to the other will be straightforward. Furthermore, the phased

approach maximises the possibility of exploiting the most recent advances in computer technology since we will not be wedded to one hardware solution from the start and will delay as long as possible each hardware upgrade acquisition.

4 PIXEL, CATALOGUE & WEB SERVERS

4.1 Pixel server

The baseline requirements for the pixel storage are high capacity and expandability to the tune of 22 Tbytes per year. Low–cost, mass storage of pixel data is a solved problem: the ESO Next Generation Archive Systems Technologies [2] employs low–cost IDE disks connected to mid–range CPUs to provide multi–Tbyte capacity. We will implement an NGAST–like solution for WSA pixel storage; however we will implement RAID level 5 to include fault tolerance against individual disk failure – NGAST apparently does not currently use RAID. The operating system will be linux and FITS data will be stored as flat files in an observation–date driven directory structure. The FITS data stored will consist of images (and also catalogue FITS binary table files) produced by the CASU processing pipeline – for more details, see the ICD (AD02). The pixel server hardware will consist of a 2.4 Tbyte file server employing a 3Ware Escalade IDE RAID controller and 12× 200 Gbyte IDE disks along with nine further cloned nodes in a rackmount unit yielding $\sim$ 22 Tbytes of storage space after RAID overheads (see Figure 1). This system is modular and will be expanded/upgraded over the next few years of operation.

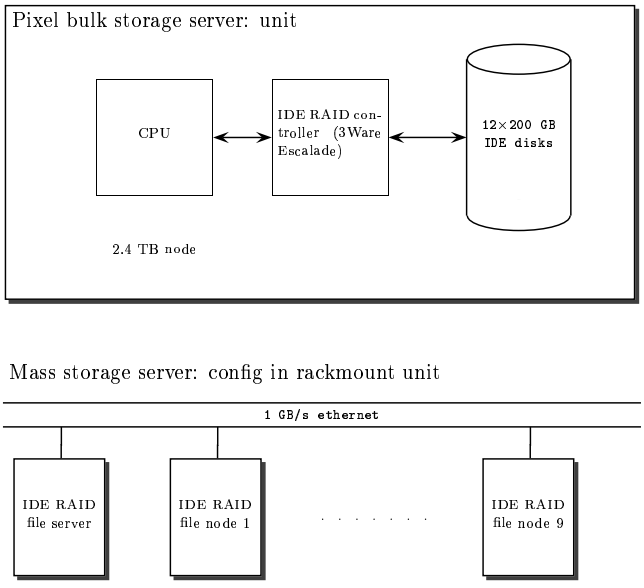

Figure 1: *The mass storage server: file server unit (above) and the rackmounted node configuration (below).*

The V2.0 pixel storage hardware solution will be the same as the V1.0 system. We will add more file servers as and when needed, phasing purchase to maximise the return per unit cost and to take advantage of any new developments in storage technology (eg. we anticipate availability of 300 Gbyte IDE hard disk drives at some point during 2004).

4.2 Catalogue servers

The baseline requirements for the catalogue hardware are:

- storage capacity of initially 2.5 Tbytes and then 2.5 Tbytes added for each year of operation;
- ability to trawl Tbyte–scale databases in a *reasonable* time – a goal of 100s is suggested but not critical; clearly a response time of many thousands of seconds limits user interaction.

Single disk performance for 100% sequential reads is typically in the range 10 to 50 Mbyte/s, depending on make, model and controller (we note in passing that the commonly held belief that SCSI interfaced disks are always faster than IDE is not based on experimental facts for 100% sequential reads [3] – it is only during random reads that the 2× faster seek time of SCSI disks produces higher IO bandwidth). For a typical single disk bandwidth of 30 Mbyte/s, trawling a 300 Gbyte table for a query on un–indexed attributes (eg. a datamining query searching for a rare type of astronomical object, the search being predicated on an unusual combination of rarely used attributes) would take $\sim$ 3 hours. Obviously, the only way around this fundamental limit is parallelisation. This could be achieved using a 'PC farm' – small numbers of disks attached to many individual CPUs – or via a RAID array attached to a single CPU, where striping across many disks allows parallelisation during IO with a consequent gain in aggregate bandwidth.

In the preliminary design phase of the WSA we suggested that the PC farm route may be a good option since as well as achieving high aggregate IO, the system automatically has at its disposal large processing power that may come in useful for advanced applications. However, the disadvantage of the PC farm is in expense and management, and it turns out that a trawl rate of < 1000 s can be achieved, at least for moderately sized tables, using inexpensive RAID technology, eg. [4]. In this study, by careful design with due regard to disk, disk controller and PCI bus bandwidth limits, aggregate IO rates of well over 300 Mbyte/s were achieved. This study used Ultra160 SCSI controllers along with SCSI disks, and matched the number of disks and their bandwidths to the measured saturation limits of the controllers; note that software striping across the disks was employed. A useful figure to come out of this and other similar studies is that the manufacturer's 'burst' transfer specification for any device (eg. 160 Mbyte/s for Ultra160 SCSI controllers) will typically fall by 25% for sustained IO rates. So, for example, Ultra160 controllers are capable of sustaining IO rates of 120 Mbyte/s – hence a maximum of 3 disks, each giving 40 Mbyte/s, were attached to each controller in [4]. The key point to note concerning optimising aggregate IO rates for a hardware system is that the saturation limits of each component in the IO chain – PCI bus, interface connection card, disk controller(s) and disk(s) – must be carefully considered and matched such that no one component limits the potential performance of the rest (the ultimate limit of a single CPU system is the CPU and PCI bus, which can typically shift data at rates of 0.5 to 1.0 Gbyte/s).

The disadvantages of the configuration in [4] as regards the WSA requirements are, for optimum performance:

- no fault tolerance is present (ie. no RAID redundancy);
- capacity is limited to that achievable via the available interface card slots to the PCI bus on the CPU motherboard, the number of disks per interface/controller an the capacity of the disks themselves.

Figure 2 shows our catalogue server design for the WSA, where we employ the IO advantages of [4] but using low cost IDE disks and RAID controllers to achieve the necessary storage capacity, fault tolerance and high aggregate IO, all at reasonable cost. The CPU will be a dual processor system employing Xeon Pentium IV 2.8 Ghz processors and a PCI–X (64 bit, 133 MHz) data bus. This design

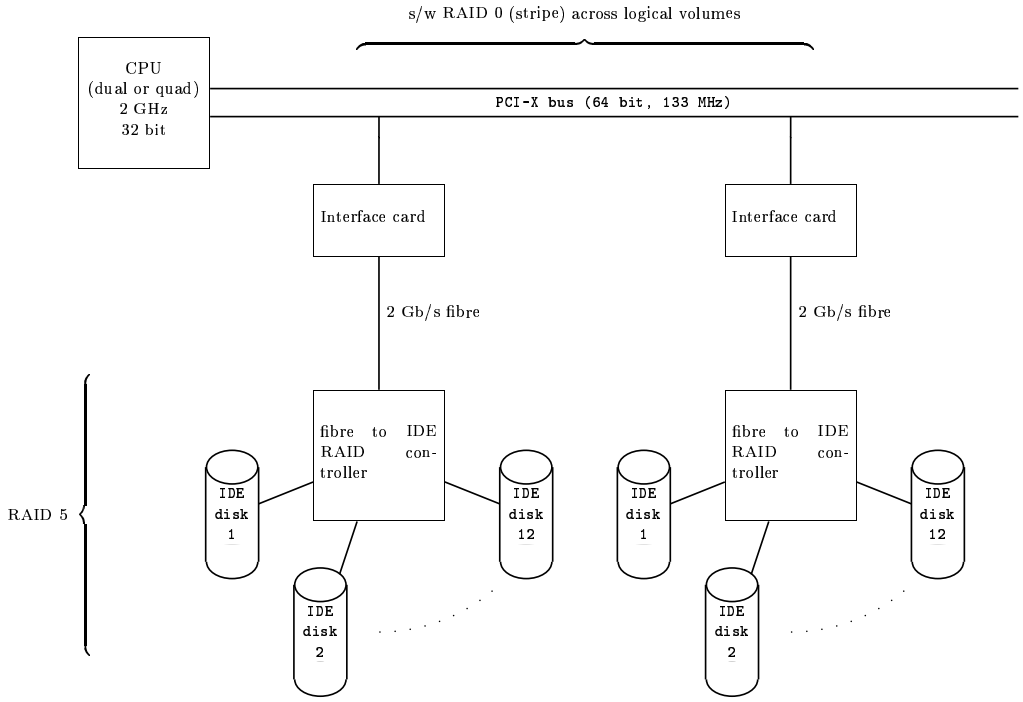

Figure 2: *The catalogue server system for the WSA*

follows testing using our own, and borrowed hardware: we have tested the fibre–to–IDE solution trawl rate performance (and also some other hardware configurations – see Appendix 8.2) for one IDE RAID array unit using real–world astronomical queries and the OS/DBMS choice we will use for the WSA (see below; again, more details are given in Appendix 8.2). The results are shown in Figure 3, where several trawl–type queries were executed per disk array configuration (all RAID level 5) using a 5 Gbyte table to ensure no misleading results from caching anywhere in the system.

The performance of the fibre–to–IDE controllers used here is not as good as might be expected given that the single IDE disks are capable of reading at sustained rates of 40 Mbyte/s; additionally the saturation performance of one controller ($\sim$ 80 Mbyte/s) is not on its own up to the requirements. This needs a little more experimentation and optimisation before specifying and ordering the V1.0 hardware. We suspect (but do not yet have experimental evidence to support this suspicion) that the drop in per–disk performance to $\sim$ 10 Mbyte/s and the saturation at $\sim$ 80 Mbyte/s are inherent to the RAID controllers, and hence this is the price to be paid for high capacity and fault–tolerant inexpensive disk arrays. In the design illustrated in Figure 2, we will use software RAID0 striping over the logical volumes presented by two RAID controllers to further parallelise the IO up to $\sim$ 150 Mbyte/s.

At the time of writing, we are borrowing Ultra320 SCSI hardware to compare the performance and cost of a configuration closer to [4] in order to inform the final V1.0 hardware decision; this comparison will include an investigation of the performance of the Ultra320 devices as hardware RAID controllers (for fault tolerance).

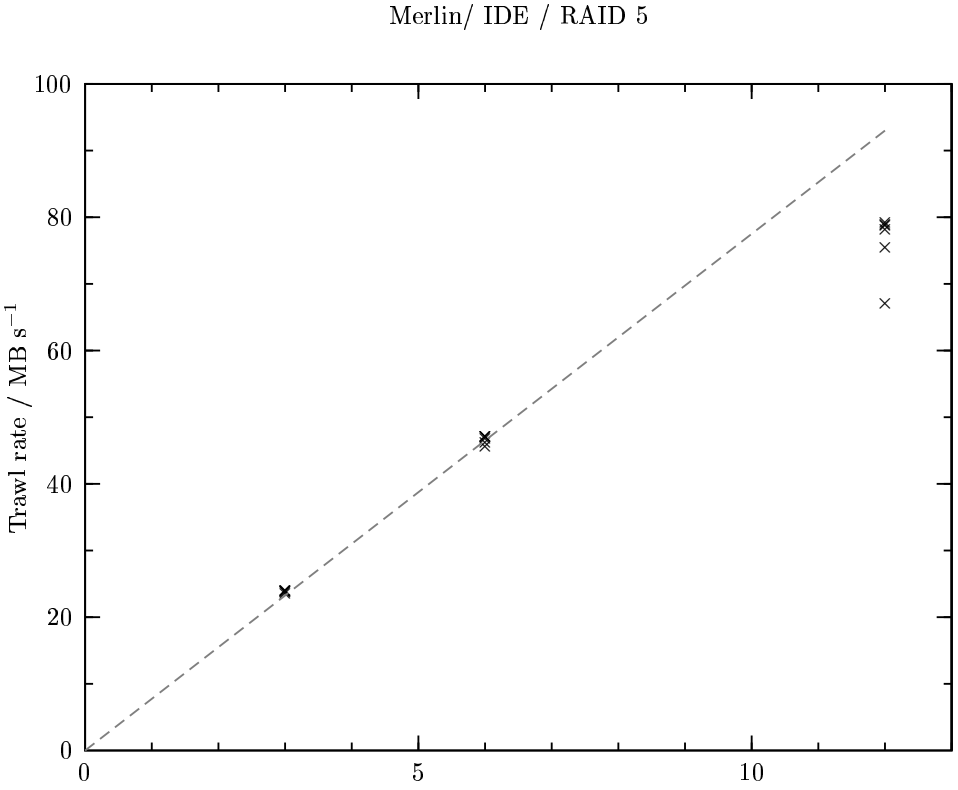

Figure 3: *Trawl results for the fibre–to–IDE catalogue server solution; the X–axis units are no. of disks in the RAID configuration. Crosses indicate individual trawl queries for a given disk array configuration; the straight line is the 3–disk array configuration linear scaling.*

4.2.1 OS/DBMS choice

The baseline requirements for the catalogue OS/DBMS choice are flexibility (eg. provision of industry–standard SQL interface, ability to cope with Tbytes of data) and ease of use. Previously, we have implemented WFAU data services with *ad hoc* flat file systems, but such solutions are not flexible or scalable to large data volumes. Recently, we have been following SDSS science archive developments in order to benefit from the experience and software developed for the SDSS archive. We have used MS Windows and SQL Server to provide a flexible archive for the '6dF' database [6]. We have designed a schema for an SQL Server implementation of our own SuperCOSMOS Sky Survey (SSS [7]) – for more details, see the Database Design Document AD04 and references therein. We have successfully ingested SSS data and exercised this database for 4×10^7 records, or 1% of the total SSS size, in the SDSS–EDR regions (eg. the tests reported in Appendix 8.2). Our experience with all of the above indicates that Windows/SQL Server is an excellent choice for the V1.0 WSA. We will prototype the V1.0 WSA hardware solution by implementing the entire $\sim$ 1 Tbyte SSS archive by the end of June 2003 to produce an externally queryable, Tbyte–scale archive for outside users to test prior to implementation of the V1.0 WSA (this service is to be known as the SuperCOSMOS Science Archive, or SSA).

4.2.2 'Load' and 'public' catalogue servers

Given our own experience of curating/serving the Tbyte–scale SSS, and on advice from our colleagues in the SDSS science archive group, we will implement a hardware design that consists of two independent SQL Servers: a 'load' server and a publicly accessible 'query' server. The main reasons for this are:

- constant (daily) updates will occur to the database, with large IO and processing overheads (eg. ingest, indexing and other curation tasks) – it is important that these do not impact the performance of the system as perceived by external users;
- more importantly, external users must see static, 'released' catalogue products that do not change minute by minute, day by day – it will be impossible to do accurate, quantitative science with a database in a constant state of flux.

Hence, the 'load' server will be used for daily curation and will be accessible only internally within WFAU, while the 'public' server will be used for user access of released catalogue data products. This of course doubles the required storage capacity for catalogue data.

4.3 Web server

Security issues will be dealt with using solutions based on our experience in setting up secure web servers and data services. We require one other server for the purpose of data service connection to the outside world. Our existing SQL Server data services are isolated from the internet (it is not recommended to expose Windows database servers directly to the internet), and our existing online services (see, for example, the user interface document, AD05) are implemented on linux web servers connecting the users to the SQL Server via Apache/tomcat. We will follow this experience in implementing the WSA. The web server will be a mid–range linux PC; the local storage requirements will not be high so a 200 Gbyte hard disk drive will suffice; however some user access–time pixel manipulation will be necessary on this server so 2 Gbytes of physical memory will be required.

4.4 V2.0 considerations

The V2.0 catalogue systems must be scalable to data volumes $\sim$ 10 Tbytes and beyond. Currently, there is some uncertainty as to the suitability of SQL Server for this; maintaining rapid response for such large data volumes requires very high aggregate IO which in turn will probably require CPU parallelisation. Oracle or DB2 (running on unix/linux) may be a more suitable choice, but this needs more investigation. We have engaged academic database researchers in the University of Edinburgh, notably Profs. Peter Buneman and Malcolm Atkinson, who are both internationally regarded authorities in the indexing of large databases. We have established contacts with Ian Carney (Oracle) and Andy Knox (IBM/DB2) for technical advice concerning their respective DB management systems, and have instigated an R&D strand to the project to address scalability issues through V2.0 and beyond. At the time of writing, meetings are scheduled for April 8th and June 30th at the National e–Science Centre in Edinburgh with these external contacts and with Jim Gray (MS research) and Alex Szalay (SDSS at Johns Hopkins University) to progress this work. Our top level plan (see the Management and Planning document) shows how this strand fits into the overall project, and shows a review milestone at the end of Q1 2004 to examine the V2.0 hardware/OS/DBMS solution proposed.

Note that our V1.0 pixel solution is modular, scalable and limited in capacity only by physical accomodation issues – these are addressed in Section 7.2.

5 EXTERNAL NETWORK CONNECTIVITY

We do not anticipate any problems in achieving the required network bandwidth to routinely transfer processed data from CASU to WFAU.

The requirement for external network connectivity is 6 Mbytes/s continuous bandwidth. All UK HEIs and data centres, including WFAU and CASU, are interconnected using the Joint Academic Network (JANET;[8]). We have tested the current, standard bandwidth between CASU and WFAU (see Appendix 8.4) and measured a bandwidth of $\sim$ 1 Mbyte/s; we note that further tests [9] between CASU and the LEDAS data centre at Leicester University have achieved $\sim$ 4 Mbyte/s while typical rates between any two JANET sites are $<$ 1 Mbyte/s.

We have also consulted with our local networking experts within ATC/IfA computing support and Edinburgh University Computing Services, and we have investigated transfer protocols and have mapped out the network between WFAU and CASU (see Appendix 8.3). Noteworthy points are as follows:

- the fundamental limit to transmission times is of course dictated by signal propagation delay (essentially light travel times) in the network links; these are well below other processing delays etc. but end–station buffers should be large enough to hold blocks of data during the 'flight time' of data in the system;
- there are 12 'hops' in the standard CASU/WFAU path; each hop introduces latency which can additionally limit available performance;
- the actual bandwidth obtained in the tests were limited by the 100 or 10 Mbit/s links in the servers at either end of the transfer chain.

The following changes will be made to the external network connectivity to achieve the required bandwidth:

- default TCP buffer sizes at each end of the network chain will be increased from the default (64 Kbyte) to 256 Kbyte in line with measured round–trip times of 15 ms (the calculation is 6 Mbyte/s $\times$ 0.015s=90 Kbyte, so 256 Kbyte leaves plenty of spare capacity);

- the WSA will be connected directly to the JANET backbone via the newly upgraded local connection provided via the Strategic Research Infrastructure Fund (known as the 'SRIF' network; see Appendix 8.3) – this bypasses local, heavily used ROE and University parts of the network by cutting out two local network hops (the expected connectivity will then be as illustrated in Figure 7 in Appendix 8.4);
- an independent WSA LAN (1 Gbit/s) with independent, internally–firewalled servers will connect the archive hardware to JANET via the SRIF network – again, this bypasses the general ROE site firewall.

The bandwidth limit will then be dictated by the CASU end connection to their RAID store. An upgrade to the server at that end will be done if found to be necessary. We are currently testing GridFTP for the network transfer from CASU (see Appendix 8.3 for brief details of several transfer protocols).

6 BACKUP

The fundamental requirement for data security is that backups are essential for *all* data. Raw data are not a WFAU concern, but we note in passing that offline raw data copies of WFCAM data will be held at JAC and CASU (ESO will also have a raw data copy). Processed pixel data and standard catalogue detection products associated with those images will be stored on spinning disk and in an offline archive at CASU; processed pixel and catalogue data will of course be held online on spinning disk at WFAU on fault–tolerant RAID5 systems (see previously). We do not expect to make offline copies of the large volume processed pixel data because in the event of data loss, the affected files will be recoverable from the CASU backup. However, our experience is that catalogue product backups are highly advisable when wishing to provide a reliable service to users; for example, our SSS data have needed to be recovered from *removable* backup media once in the past, avoiding on online service interuption of several months. Hence we will use the latest high capacity system for removable media backups: 'Ultrium' LTO–2 tape. There are several features of these systems that make them ideal for WSA catalogue backups:

- each tape has a 200 Gbyte native data capacity;
- tape 'library' configurations are available with 30 slots (and up to 6 drives per library) and hence a one–off backup capacity of 6 Tbyte;
- the transfer rate is $\sim$ 100 Gbyte/hour, enabling overnight backup of $\sim$ 1 Tbyte (or weekly, overweekend backups of many Tbytes);
- there is a clear upgrade path with these units: LTO–3 and 4 upgrades are in development [10], each of which will double capacity and speed (so LTO–4 is expected to have 800 Gbyte native capacity with a transfer rate of 400 Gbyte/hour);
- drives/tapes will be upgradable in existing library hardware – so to upgrade from LTO–2 to 3 one keeps the existing library and simply changes the drive(s) and tapes.

Hence, we anticipate that LTO–2 will easily keep pace with backup capacity requirements over the next few years as data accumulate.

We note that per Tbyte, tape is still the cheapest, most flexible, and most secure method of making removable media data copies.

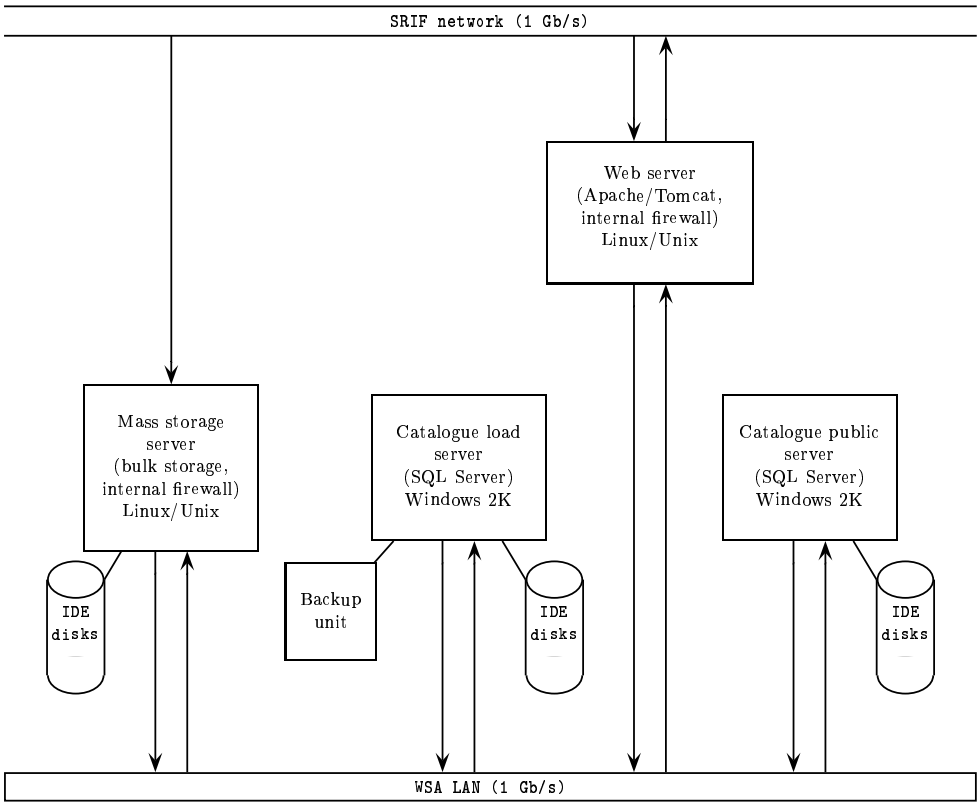

Figure 4: *Overall WSA hardware architecture: servers, LAN and connection to JANET via the SRIF connection.*

7 LOCAL INFRASTRUCTURE

7.1 Local Area Network

We will isolate the WSA hardware from the site–wide LAN so as not to impact general on–site network performance with the heavy WFCAM data transfer load expected. To this end, a small, 1 Gbit/s LAN specifically for the WSA will be used. The mass storage pixel server and web server will be internally firewalled and connect to the SRIF connection and this WSA LAN. The resulting overall picture of the WSA hardware is shown in Figure 4.

7.2 Accommodation

Our present data service hardware are accomodated in a secure, air conditioned area protected by automatic fire extinguishing equipment. Sufficent power and space for the WSA V1.0 hardware are available in this same room; with some rearrangement we anticipate that additional V2.0 equipment will also be able to be accomodated in this room for the next 2 years. After this, we expect Super-

COSMOS plate scanning operations to be coming to an end, and anticipate being able to refurbish the existing computer area (which has the necessary air conditioning, automatic fire extinguishing, power and network infrastructure already in place) for the purposes of WSA hardware.

7.3 System management

Effort for system set–up and management is available within the existing staff effort allocated to the WSA project (see the management and planning document, AD06). A small amount of additional computing support is available as part of the general allocation available to WFAU from the ATC/IfA computing support team. This includes staff experience in networking and management of both linux and Windows systems.

References

[1] Astronomical Wide–field Imaging System for Europe;
http://www.astro-wise.org/

[2] ESO Next Generation Archive Systems Technologies;
http://archive.eso.org/NGAST

[3] Performance comparison of IDE and SCSI disks;
http://www.cs.virginia.edu/~bsw9d/papers/ide_scsi.pdf

[4] Data Mining the SDSS SkyServer Database, Microsoft technical report MSR–TR–2002–01 (January 2002); ftp://ftp.research.microsoft.com/pub/tr/tr-2002-01.pdf

[5] The Sloan Digital Sky Survey at Johns Hopkins University; http://www.sdss.jhu.edu/

[6] The 6dF Galaxy Survey database; http://www-wfau.roe.ac.uk/6dFGS

[7] The SuperCOSMOS Sky Survey; http://www-wfau.roe.ac.uk/sss

[8] http://www.ja.net/topology/

[9] M.J. Irwin, private communication

[10] Eclipse Computing, private communication

[11] WFCAM Science Archive Science Requirements Analysis Document;
http://www.roe.ac.uk/~nch/wfcam/srd/wsasrd/wsasrd.html

[12] WFCAM Science Archive interface control document;
http://www.roe.ac.uk/~nch/wfcam/VDF-WFA-WSA-004-I1/VDF-WFA-WSA-004-I1.html

[13] WFCAM Science Archive database design document;
http://www.roe.ac.uk/~nch/wfcam/VDF-WFA-WSA-007-I1/VDF-WFA-WSA-007-I1.html

[14] WFCAM management and planning document;
http://www.roe.ac.uk/~nch/wfcam/VDF-WFA-WSA-002-I1/VDF-WFA-WSA-002-I1.html

[15] The UKIDSS Proposal; http://www.ukidss.org/sciencecase/sciencecase.html

[16] 20 Queries for the SuperCOSMOS Science Archive
http://www.roe.ac.uk/~nch/wfcam/misc/20queries.sql

8 APPENDICES

8.1 Data volumes and rates

The following assumptions can be made about WFCAM; these are unlikely to change:

- the maximum data rate from the DAS is one 4–device 'footprint' per 10s;
- the DAS will output 4 bytes per pixel;
- the DAS will co–add sky limited sub–exposures when a single demand exposure (specified for observing efficiency reasons for example) would saturate in the sky (eg. 40s demand integration at K);

Here, the 'DAS' (≡Data Acquisition System) consists of the device itself, the device controller, and ADC and a PC system that processes the 16 bit/pixel reads from the ADC. There will be four such subsystems operating in parallel, one per WFCAM device. The PC will coadd individual 16–bit reads, or process non–destructive reads etc., and for generality will always output 32 bit values ≡4 bytes/pixel (note: these PCs should *not* be confused with those that will operate the summit pipeline). We further assume that the individual sky–limited sub–exposures will not be kept. For the purposes of estimating both the maximum and typical data volumes/rates, the following further assumptions can be made:

- online archive will not utilise any 'lossy' compression to reduce pixel storage requirements;
- UKIDSS [15] ($\sim$ 700 night combined science programme) is typical of the science that will be undertaken with WFCAM;
- assume 2×2 microstepping and WFCAM project scientist efficiency estimates for the purposes of time–average numbers (eg. the UKIDSS proposal);
- the PEAK data rate is that delivered by the DAS maximum and using a 2×2 microstep mode with efficiency 0.65. An interlaced frame will be produced in $\sim$ 1 minute comprising 4 detectors $\times$ 2048 $\times$ 2048 pixels $\times$ 4 bytes per pixel $\times$ 2×2 microsteps = 270 Mbyte or 230 Gbyte per perfect 14 hour winter night;
- Typical time–averaged data rate can be estimated by averaging over UKIDSS component surveys, assuming an average 10 hour night and scaling the data rate in the previous item by the relative survey efficiencies:

LAS	183.4 nights	160 Gbyte per night =	29.3 Tbyte
GPS	130.2	92	12.0
GCS	58.8	160	9.4
DXS	123.9	52	6.4
UDS	130.2	52	10.8
Totals:	$\sim$ 700		$\sim$ 70.0

  or an average over the science programme of $\sim$ 100 Gbyte per full night.

- data volumes/rates will include a 10% overhead on pixel data alone, where appropriate, for the purposes of allowing for derived object catalogues, housekeeping and DBMS

This last item needs closer inspection. AD04 and references therein detail the baseline set of parameters per detected object from CASU standard pipeline processing. Assuming 4 bytes per parameter (they will be mainly single precision floating point numbers) this is 80 parameters $\times$ 4 bytes = 320 bytes per detection. Further, for the purposes of list–driven co–located photometry (see the SRAD; ie. given a detection in one passband, what are the object parameters at fixed positions using fixed apertures and profiles in all other passbands) this value should be scaled appropriately for ~ 4 UKIDSS passbands. So, to order of magnitude, the catalogue record size is $\sim 10^3$ bytes per detected object. Now, the number of detected objects per frame will vary enormously. For example, in the UKIDSS GPS, towards the Galactic centre the surface density of sources is likely to be $> 10^6$ per sq. deg. (or $\sim 10^{-2}$ objects per pixel) while in the lowest surface density regions of the LAS this is likely to drop to $\sim 10^3$ per sq. deg. (or $\sim 10^{-5}$ objects per pixel). If we assume a typical surface density of sources as being $\sim 10^4$ per sq. deg., or $\sim 10^{-4}$ objects per pixel, then for a given amount of pixel data the object catalogue overhead is

$$\frac{10^3 \mathrm{bytes/obj} \times 10^{-4} \mathrm{obj/pix}}{4 \mathrm{bytes/pix}} \times 100 \approx 3\%.$$

Allowing for housekeeping, other ancilliary data and DBMS overheads, a figure of 10% overhead on pixel data does indeed seem reasonable.

An estimate of the yearly rate can be made as follows. Nights per year are likely to be 365$\times$ 80% UK time on UKIRT $\times f_{\mathrm{WFCAM}}$, the fraction of all UK time given over to WFCAM. Assume 110 Gbytes per night average, and for the likely range assume $0.6 < f_{\mathrm{WFCAM}} < 0.8$. Then, the average yearly data accumulation rate will be between 19 and 26 Tbytes.

In summary, data flow for the WSA will be:

- Ingest: $\sim$ 200/100 Gbyte per day (peak/average) for $\sim$ 200 days per year;
- Accumulation of data: $\sim$ 22 Tbytes per year;
- Accumulation of catalogue, housekeeping and other ancilliary data: $\sim$ 2 Tbytes per year.

An estimate of the final pixel storage requirement for UKIDSS at least is straightforward: assuming 4 bytes per pixel and 2×2 microstepping (ie. 0.2 arcsec pixels); the areas of the LAS, GPS, GCS, are respectively 4000 sq. deg. $\times 5$ filters; 1800×5; 1600×4 (stacked pixel data for the DXS and UDS are negligible for these purposes). This adds up to $\sim$ 50 Tbytes; the final UKIDSS object catalogues and associated data will be $\sim$ 5 Tbytes.

The uncertainties above (eg. detected objects per pixel; the amount of confidence array information needed to be stored, etc.) should not prevent progress on hardware design and acquisition, since storage for the final data volume does not have to be purchased up front. Provided sufficient storage is acquired for the first year of operation, it will become clearer during that time what the precise long term requirements are. In any case, the lifetime of the WSA project is significantly longer than the typical timescale of leaps in computer hardware design, so it should be expected that the initial hardware solution will not be the final one, and a phased approach (as is required from the science exploitation point of view; see the SRD) is implied.

8.2 Catalogue system performance tests

Given the large range of possible hardware solutions and configurations that would be implemented for the WSA, we have been conducting a test programme using our existing, and also loaned hardware, to determine the best compromise between performance, cost, scalability and complexity. The tests have taken the form of employing 20 real–world astronomical queries [16] developed for the SQL Server implementation of the SuperCOSMOS Sky Survey, or SSA. These queries include several examples

Hardware configuration name	Average Trawl rate (Mbyte/s)	Disk interface	No. of disks	Disk array configuration	Note
Grendel01	27	SCSI (Ultra160)	3	SW RAID0	800 MHz PII
Lancelot	27	SCSI (Ultra160)	3	RAID5	Xeon quad proc (4x1.6 GHz)
merlin1	24	fibre–to–IDE	3	HW RAID5	AMD 2200 XP proc (1.8 GHz)
merlin2	77	"	12	"	"
merlin4	46	"	6	"	"
merlin5	75	"	12	RAID1+0	"
merlin6	53	fibre–to–fibre	6	RAID3	"
merlin7	60	"	6	SW RAID0	"
merlin8	47	"	5	RAID5	"

Table 1: *Some trawl benchmarks for various hardware configurations; results for 'merlin' 1,2 & 4 are plotted in Figure 3 to show linear scale–up with disk array no. and also saturation limits for the particular controller in question.*

pertinent to expected usages of the WSA, eg. joint queries with the SDSS; however for the purposes of trawl benchmarking we have used a subset of 6 of the 20 queries that trawl the multi–epoch, multi–colour merged SSA catalogue corresponding to the SDSS–EDR regions. This catalogue is 4.87 Gbyte in size. The performance figures are summarised in Table 1.

The important point to note here is that the use of hardware RAID controllers appears to compromise the per–disk performance of the system: for example, the IDE disks used in configurations 'merlin' 1 to 5 have been measured to be individually capable of sustained transfer rates of $\sim$ 40 Mbyte/s whereas the 3 disk RAID5 set configured in 'merlin1', for example, delivers only 24 Mbyte/s aggregate IO.

At the time of writing, our test programme is not quite finished. We are currently benchmarking Ultra320 SCSI controllers under otherwise identical conditions in order to compare the cost/performance trade–off before we finally place an order for the first V1.0 catalogue server.

8.3 Networking

8.3.1 Morphology

Network morphology is illustrated in Figure 5. The circle in the top left shows the location of the current WFAU data servers on the network: cosaxp6 (SSS & 6dF web server) and grendel12 (current WFAU/AstroGrid linux/apache web server) is on the 100 Mbit/s LAN, wiglaf (Sloan EDR mirror) is on the 10 Mbit/s DMZ and grendel10 is being set up on the SRIF network (as srif112) to test its use (grendel10 only has a 100Mbit/s network card at the moment). The second router on the LAN is required to change from 100Mbit/s copper to Gbit/s fibre.

Network in and around the University is illustrated in the top right of Figure 5. The Strategic Research Infrastructure Fund upgrade to the Edinburgh network infrastructure is depicted in the softbox labelled SRIF. The core of the SRIF network can be thought of for present purposes as a set of four 1 Gbit/s fibres running between a pair of routers. One of these is connected to the University network, and the other is connected to a router (at the university King's Buildings) to which the ROE LAN attaches. That router is then connected to a router to which EaStMAN attaches, and that router, in turn, connects to the Edinburgh BAR (Backbone Access Router), which connects to the JANET backbone. All the connections shown here are 1 Gbit/s. A new Gbit/s link has been agreed from the SRIF router to which the ROE SRIF link attaches straight to the BAR, so that SRIF traffic can be sent to JANET on a separate route from UoE and EaStMAN traffic.

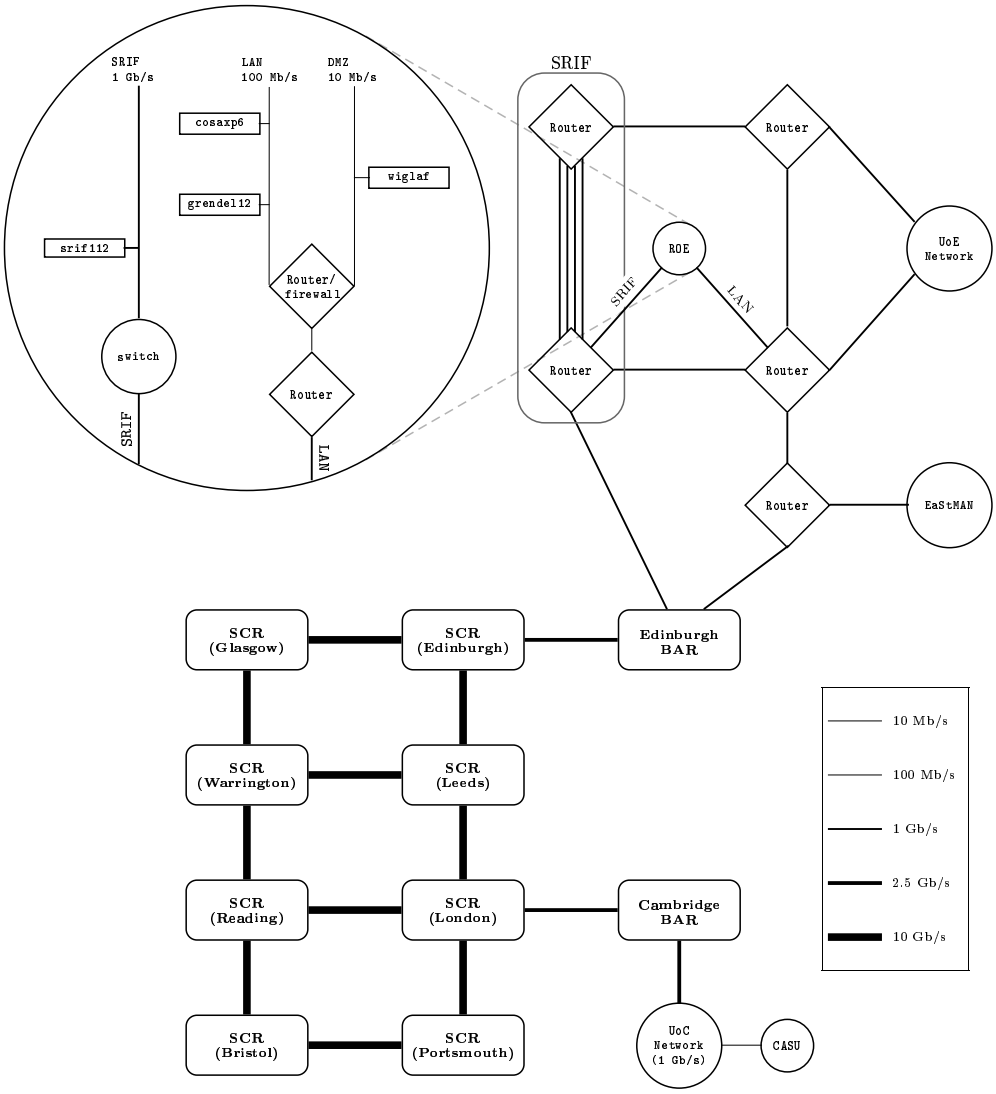

Figure 5: *Schematic of WFAU/CASU network connectivity with detail at the WFAU end, location of relevant new SRIF connections, and including the JANET backbone. For a description, see the text.*

The rest of Figure 5 shows the basic morphology of JANET. Traffic from the Edinburgh BAR gets to Cambridge via the SuperJANET Core Routers (SCR) in Edinburgh, Leeds and London, and there is a 2.5Gbit/s link from the London SCR to the Cambridge BAR, which attaches to EastNet and the Cambridge University network (some of the links between SCRs are being upgraded to 10Gbit/s, and, indeed, our test results (see below) suggest that the Leeds–London link of our route to CASU is running at 10Gbit/s.

8.3.2 Transfer protocols

Most data transfer over the Internet proceeds through the use of TCP/IP, which is a suite of protocols.

IP (Internet Protocol) is a protocol for sending packets of data. It includes no notification of arrival, nor does it guarantee the order in which packets will arrive or when they will arrive.

TCP (Transmission Control Protocol) sits on top of IP, usually implemented in the kernel of Unix machines. Its operation can be best understood through a fiction in which there is a duplexed data stream (send & receive) running between applications sitting on the machines at each end of a data transfer: it is conventional to refer to one of these connections as 'a TCP'. In this fiction, data either arrives in the correct order or the connection is broken. The sender numbers packets and requires acknowledgement of receipt of the packets. The sender buffers data packets until it receives an acknowledgement that the particular packet has been received, or until a time–out occurs due to a broken connection. So, clearly, if the sender's buffer is too small, it may fill up before the first acknowledgement is received, at which point it has to stop sending more packets. The acknowledgement is in the form of an indication of how much space is available in the buffer at the receiving end, and the sender acts conservatively on that information, so that, if the buffer space at the receiving end is decreasing, it slows down the speed at which it despatches more packets. So, a larger buffer at the receiving end is desirable, too (the obvious inference would be to try to have as large a buffer at either end as possible, and there would seem to be no reason to limit the buffer sizes, were it not for the fact that doing so would tie up possibly unnecessarily large amounts on memory in the two end–station machines). The receive buffer will be filled before the first acknowledgement is received by the sender if its size is less than the product of the data flow rate into the connection (assumed constant) and the round–trip delay time along the connection. This is called the Bandwidth–Delay Product (BDP), and it is a very useful quantity in analysing network performance.

The fundamental limit in the delay is the light–travel time between the end–stations, but in realistic systems the overall delay is significantly above this limiting value, due to processing delays, etc. As the distance between the end–stations increases, the delay time increases, more data is in flight and the receive buffer must be larger to cope with it. For example, given the delay time measured in tests between Edinburgh and Glasgow, 256KB buffers would be required for a Gbit/s link. Tests with a large-buffered system have attained transfer rates something like 450Mbit/s for this dedicated connection, which is close to the speeds of the internal buses in the machines, showing that, with correct configuration, the limiting factor can be the hardware at each end.

The default TCP buffer size is 64kB and the round trip light travel time between Edinburgh and Cambridge (1000km round trip) is about 3ms. This would suggest that the default buffer size could support a transfer rate of 6.4MB/s, which would be sufficient for WFCAM. However, tests record round trip travel times of 15ms, so the default buffer size cannot handle the sustained 5MB/s needed for WFCAM.

FTP is an application that runs on two TCP connections: it uses one for controls and another to send the data. However, a new application known as GridFTP can open parallel connections, and so can attain a higher aggregate bandwidth, by striping data transfer across them.

For completeness, UDP (User Datagram Protocol) is another protocol which sits on top of IP, but which has a much less functionality than TCP. In UDP, there are no acknowledgements and no error

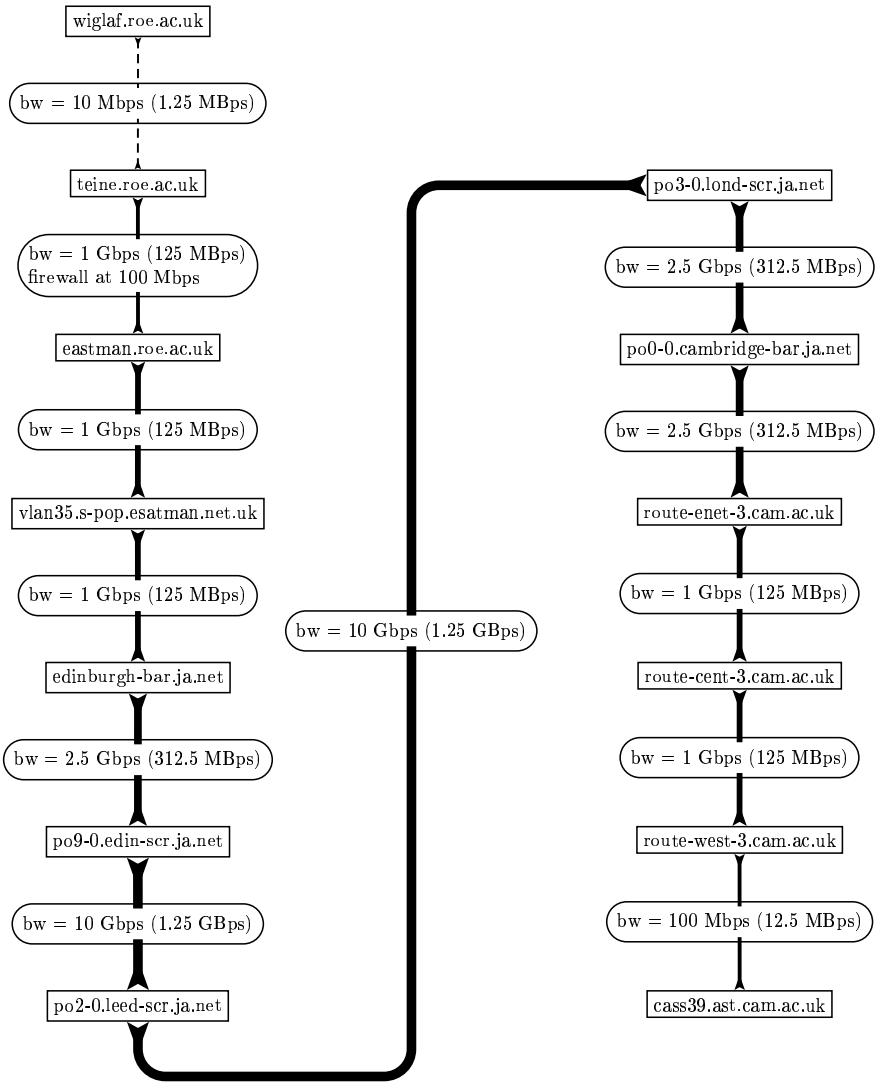

Figure 6: *The actual network bandwidth between ROE and CASU as of March 2003, with bandwidth specifications for each connection in the web.*

recovery; in essence, UDP just squirts lots of data into a connection, and whatever application is on the other end has to deal with whatever arrives. This is fine for things like video streams, where a lost frame is of no worth once the subsequent frame has been displayed, but, for something like FTP, where every packet is needed, the use of UDP necessitates that the application software handle everything that is otherwise done by TCP. This puts a lot of extra burden on the application code, so we do not consider UDP as a viable solution for transfer of WFCAM data.

8.4 Network tests

Figure 6 shows the specified network bandwith between ROE and CASU at each hop in the chain represented in Figure 5; Figure 7 shows a similar picture when connecting to the JANET backbone via the SRIF network.

To test the data transfer and measure real transfer rates to compare with those specified in Figures 6 and 7, transfer of data was tracked with `traceroute`. The program `pchar` was used to measure

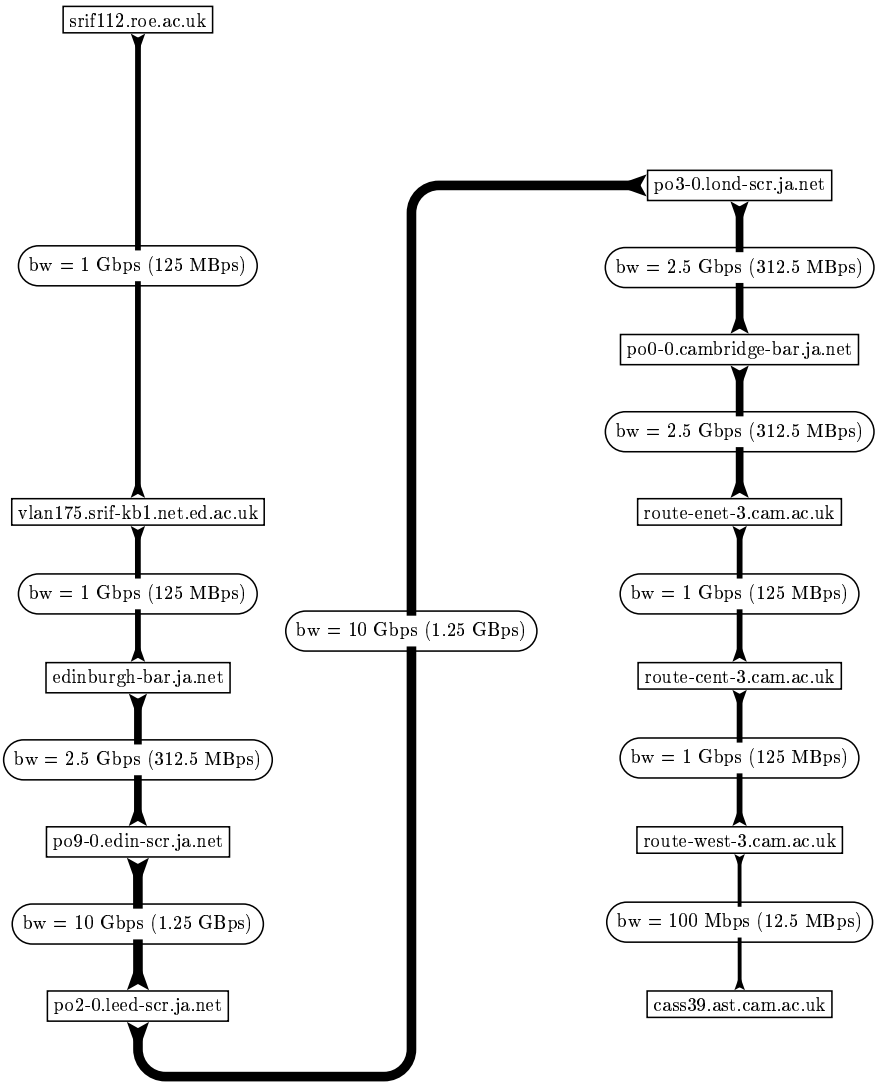

Figure 7: *The actual network bandwidth between ROE and CASU as of March 2003, with bandwidth specifications for each connection in the web.*

Host	Bandwidth (Mbyte/s)		Comment
	Quoted (max)	measured	
wiglaf	1.25	1.2	WFAU linux server
teine	125	4.9	ROE firewall
eastman	125	6.0	
vlan35	125	N/A	
edinburgh-bar	313	172	JANET backbone router
po9-0.edin-scr	1250	75	
po2-0.leed-scr	1250	N/A	
po3-0.lond-scr	313	30.6	
po0-0.cambridge-bar	313	11.8	
route-enet-3	125	106	
route-cent-3	125	263	
route-west-3	12.5	5.8	

Table 2: *'pchar' test results of the default connectivity between CASU and WFAU at the time of writing (March 2003). Using the SRIF connection bypasses hosts teine, eastman anmd vlan35, giving a 125 Mbyte/s bandwidth in a single hop to edinburgh-bar (see Figure 7).*

the characteristics of the network path between the WFAU host wiglaf and cass39 (the Cambridge FTP server). It is an independently–written reimplementation of the `pathchar` utility, using similar algorithms. Both programs measure network throughput and round-trip time by sending varing–sized UDP (User Datagram Protocol) packets into the network and waiting for ICMP (Internet Control Message Protocol) messages in response. Table 2 summarises the results. It should be noted that the measured figures can be quite inaccurate as only small packets of data are being sent in these tests; further, as pointed out previously it would appear that the routers at the Cambridge end of JANET have been upgraded since the specifications were obtained. Anyway, the main point to note here is that there is a large available bandwidth between the Cambridge and Edinburgh BARs, and that a sustained transfer rate of 5 Mbyte/s can be achieved by upgrading connections at either end as detailed in the main text.

9 ACRONYMS & ABBREVIATIONS

ADnn : Applicable Document No nn
BAR : Backbone Access Router
CASU : Cambridge Astronomical Survey Unit
DBMS : Database Management System
DMZ : De-militarised Zone
ICMP : Internet Control Message Protocol
IDE : Integrated Device Electronics
JANET : Joint Academic Network
LAN : Local Area Network
LEDAS : Leicester Data Archive Service
OS : Operating System
RAID : Redundant Array of Inexpensive Disks
SCR : SuperJANET Core Router
SCSI : Small Computer System Interconnect
SRAD : Science Requirements Analysis Document
SRIF : Science Research Investment Fund

UDP : User Datagram Protocol
VISTA: Visible and Infrared Survey Telescope for Astronomy
WFAU : Wide Field Astronomy Unit (Edinburgh)

10 APPLICABLE DOCUMENTS

AD01	Science Requirements Analysis Document [11]	VDF-WFA-WSA-002 Issue: 1.3 (20/03/03)
AD02	WSA Interface Control Document [12]	VDF-WFA-WSA-004 Issue: 1.0 (2/04/03)
AD04	WSA Database Design Document [13]	VDF-WFA-WSA-007 Issue: 1.0 (2/04/03)
AD06	WSA Management and Planning Document [14]	VDF-WFA-WSA-003 Issue: 1.0 (2/04/03)

11 CHANGE RECORD

Issue	**Date**	**Section(s) Affected**	**Description of Change/Change Request Reference/Remarks**
Draft 1	07/03/03	All	New document
Draft 2	14/03/03		
Draft 3	17/03/03	All	Rearranged
Draft 4	28/03/03	All	Rewritten
1.0	02/04/03	Minor changes	First issue (for CDR)

12 NOTIFICATION LIST

The following people should be notified by email whenever a new version of this document has been issued:

WFAU: P Williams, N Hambly
CASU: M Irwin, J Lewis
QMUL: J Emerson
ATC: M. Stewart
JAC: A. Adamson
UKIDSS: S. Warren, A. Lawrence

__oOo__