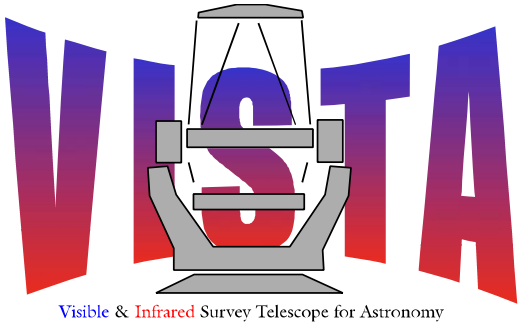

Data Flow System

Document Title: **Cambridge Reduction and Analysis Pipeline**

Document Number: **VIS-SPE-IOA-20000-0017**

Issue: **1.1**

Date: **2007-05-08**

Authors: **Jim Lewis**
Peter Bunclark
Mike Irwin (CASU)

Change Record

Issue	Date	Sections Affected	Reason/Initiation/Documents/Remarks
1.0	2006-09-28	All	New Document
1.1	2007-05-08	several	Immediate updates arising from the UK Review.

Contents

Change Record 2
1 Introduction 4
1.1 The Scope Of This Document 5
1.2 Applicable Documents 6
1.3 Reference Documents 6
1.4 Abbreviations and Acronyms 7
1.5 Glossary 7
2 The UK User Requirements 8
2.1 General requirements 8
2.2 Astrometric Requirements 9
2.3 Photometric Requirements 9
2.4 Tiling, Stacking, Microstepping 9
2.5 Variable Objects 11
2.6 Object Catalogues 11
3 Pipeline Processing 13
3.1 Overview Of Standard Processing Steps 13
3.2 Extra Processing 14
3.2.1 Background Correction 14
3.2.1.1 Sky Glow, Thermal Emission and Scattered Light 14
3.2.1.2 Fringes 15
3.2.1.3 Stripes 16
3.2.2 Tiling/Resampling 16
3.3 The RECIPE Keyword 17
4 Cambridge Pipeline Infrastructure 18
4.1 Data Transport 18
4.2 Data Ingestion 18
4.3 Triage 19
4.4 Software 20
4.5 Hardware 21
4.5.1 Processing Power 21
4.5.2 Disc Storage 22
4.5.3 Off-line archive 22
4.6 Human Resources 22
5 Risk Register 23
6 Schedule 23

1 Introduction

The Visible and Infrared Survey Telescope for Astronomy (VISTA) is a new 4-metre telescope designed specifically for imaging survey work at visible and near-infrared wavelengths. VISTA infrared camera (VIRCAM) has a field of view of 1.5^{o} in diameter with a pawprint that will cover 0.59 degree^2 in ZYJHK$_s$ passbands, using a 4×4 array of 2k×2k non-buttable Raytheon detectors with ~0.34″ pixels resulting in a 269MB FITS file per exposure. The camera is expected to produce a median of 210GB a night with realistic high volume night producing 630GB.

The VDFS has to cater to two different sets of user requirements. The first are the requirements set out by ESO and have all been thoroughly reviewed and passed by them. The UK community has an additional set of user requirements which concentrate more on the final quality of the science data products that come from the VDFS. As a result VIRCAM data will pass through at least three pipelines from the time of observation:

- **Paranal**: A causal "QC-1" (Quality-Control) pipeline will run at the telescope in near real time. It will use purpose-written recipes to compute a range of QC parameters which will enable early assessment of the quality of the data. Calibration data (such as flat fields) will be drawn from a library which will be updated only infrequently. Because of the causal nature of pipeline execution, each recipe can only make use of data taken up to that point.
- **Garching**: a calibration and quality control pipeline to be run at ESO headquarters, to remove instrumental signatures, produce catalogues and enable astrometric and photometric calibration in addition to extra quality control checks. This will use the same software modules as both the Paranal and UK pipelines.
- **UK**: The UK pipeline will be responsible for producing most of the science-grade reductions of VIRCAM data. Many of the algorithms used in the QC pipelines will be re-employed, the difference being that a more flexible approach can be taken in the production of master calibration data. Processing steps that are not possible in the QC pipelines will be run in the UK, thus improving the quality of the data products. Only calibrated science products will be released.

The UK processing will adopt the same philosophy as the WFCAM pipeline and will be based around a standard pipeline to remove instrumental signature, generate astrometric and photometric calibration, provide assorted quality control measures, and produce a standard set of science products.

In essence the standard pipeline in Cambridge will be a superset of the Garching calibration pipeline such that if the same processing sequence is used along with the same master calibration files, identical output products will result. It will operate on the Observing Blocks (OBs) taken during each night to produce the final individual

instrument signature-free images, including mosaiced tiles, and to derive astronomical object catalogues from them. The object catalogues will be used to generate the astrometric and photometric calibration, again in a similar manner to that developed for WFCAM.

A further processing stage will produce detailed estimates of the Point Spread Function (PSF) and its variation as a function of position within the array. This information will then be used to drive the PSF fitting software which adds extra entries to the catalogues derived from the PSF fitting.

The UK pipeline will be implemented to keep up readily with the expected VIRCAM data rate (which, given the estimate of 210 Gbytes/night, is comparable to WFCAM). Hence the availability of science products will be mainly driven by the frequency of ESO data delivery to CASU and the need to control master calibration updates. We anticipate that, as for WFCAM, a steady-state processing schedule that delivers science products roughly one month after the data are taken is achievable.

Data products will be made securely available for WFAU to transfer to Edinburgh from Cambridge directly via the Internet (or private UKLight circuit) with an end-to-end minimum bandwidth of 1 Gbits/s, again in a similar manner to that developed for WFCAM.

Asynchronous processing operations such as time-series generation (and analysis), deep stacking and general-purpose list-driven photometry occur more naturally at the archive end and VDFS will offer a choice of methods to do this, including use of software developed by CASU.

1.1 The Scope Of This Document

The purpose of the original paper was to serve as an overview of the UK community's requirements on the VDFS and how CASU will implement its part of them. There is also a discussion of the operational considerations of producing science data products in the UK from the flow of raw VIRCAM data. In Section 2 we give a list of the requirements set out in the *UK VISTA User Requirements* document [AD3]. If a requirement falls into CASU's work area, then a response from CASU is also included. Section 3 gives a brief description of the processing steps required to reduce VIRCAM data. It then elaborates on how the UK pipeline will improve the processing done at the summit and in Garching. The pipeline infrastructure is discussed in chapter 4. This includes descriptions of the data transport and ingestion methods, methods for detecting bad data, software and hardware infrastructure, and finally staff effort. Chapter 5 outlines the risk register for the project.

This version includes feedback from the review, and future versions may encompass revisions to the User Requirements.

1.2 Applicable Documents

[AD1] *VISTA Infra Red Camera DFS Calibration Plan*, VIS-SPE-IOA-20000-00002, issue 1.3, 2005-12-25. This was written for the ESO reviews and largely concentrates on the calibration of VIRCAM data in the summit and Garching pipelines. It describes the type of data that will be taken at the telescope and how it will be used to calibrate the science data.

[AD2] *VISTA Data Reduction Library Design,* VLT-SPE-IOA-20000-0010, issue 1.5, 2006-09-21. This was also written for the ESO reviews. The data reduction described in this documentation stops short of the extra steps specified by the UK requirements. However as the basic steps for the data processing are almost the same for all three pipelines, this provides a good description of the basic reduction to pawprint level.

[AD3] *UK VISTA User Requirements,* VDF-SPE-IOA-000090001 issue 3.0, 2005-07-05. This sets out the science requirements on the UK pipeline.

[AD4] Hodgkin, S.T., *Calibration Of WFCAM Data On 2MASS,* (2006). This gives a more detailed account of how photometric calibration has been done for WFCAM. Similar methods will be employed for VIRCAM.

[AD5] Irwin, M.J., *Pipeline Processing of Wide-Field Near Infrared Data From WFCAM,* (2006), MNRAS, in prep. This gives an in depth description of the data reduction methods used for WFCAM. The same underpinning algorithms will be used for VIRCAM data; hence it can be read in conjunction with [AD2] to understand the methods that will be used.

[AD6] Irwin, M.J. *et. al.,VISTA Data Flow System: Pipeline Processing for WFCAM and VISTA* in Proceedings of the SPIE, vol. 5493, pp 411-422 (2004). This presents an overview of the CASU pipeline development and can be retrieved at: http://casu.ast.cam.ac.uk/publications/bib/Irwin2004

[AD7] *Web-based Pipeline and Survey Progress Monitoring,* Gonzalez-Solares, E and Riello, M. This is an explanation of the survey and pipeline monitoring tools that will be made available for VISTA.

[AD8] *VDFS Pipeline/Science Archive Interface Control Document,* Hambly, N., Bunclark, P.S., Irwin, M.J., Lewis, J.R., September 2006. This specifies the detailed content and format of the FITS files used in data transfers.

[AD9] *VISTA UK Data Flow System Review Panel Report,* 2006-11-28.

1.3 Reference Documents

[RD 1] *Definition of the Flexible Image Transport System (FITS),* NOST 100-2.0. An essential document if you need to know more about FITS.

[RD 2] *The FITS image extension*, Ponz et al, Astron. Astrophys. Suppl. Ser. **105**, 53-55, 1994. A description of the multi-extension FITS which is used for all image files in the VDFS.

[RD 3] *Representations of celestial coordinates in FITS*, Calabretta & Griesen, A&A, **395**, 1077, 2002. The definitive paper on celestial world coordinate systems in FITS images.

[RD 4] *Common Pipeline Library User Manual,* VLT-MAN-ESO-19500-2720, issue 2.0.1, 2005-04-14. A ‘readers digest’ introduction to ESO’s data reduction infrastructure (CPL).

[RD 5] *Common Pipeline Library Reference Manual,* VLT-MAN-ESO-19500-2721, issue 2.0, 2005-04-08. The full reference manual to CPL.

[RD 6] Castro, S., & Zampieri, S., *Estimate of the Number of CPU Nodes to Execute the VIRCAM Pipeline in Paranal,* 2006-07-28.

1.4 Abbreviations and Acronyms

2MASS	2 Micron All Sky Survey
CASU	Cambridge Astronomical Survey Unit
CPL	Common Pipeline Library
FITS	Flexible Image Transport System
QC	Quality Control
VDFS	VISTA Data Flow System
VIRCAM	VISTA Infra Red Camera
VISTA	Visible and Infrared Survey Telescope for Astronomy
WCS	World Coordinate System
WFAU	Wide Field Astronomy Unit (Edinburgh)
WFCAM	Wide Field Camera (on UKIRT)

1.5 Glossary

Jitter (pattern) A pattern of exposure**s** at positions each shifted by a small movement (<30 arcsec) from the reference position. Unlike a microstep the non-integral part of the shifts is any fractional number of pixels. Each position of a jitter pattern can contain a microstep pattern.

Microstep (pattern) A pattern of exposures at positions each shifted by a very small movement (<3 arcsec) from the reference position. Unlike a jitter the non-integral part of the shifts are exact fractions of a pixel, which allows the pixels in the series to be interlaced in an effort to increase resolution. A microstep pattern can be contained within each position of a jitter pattern.

OB Observation Block

Pawprint 16 non-contiguous images of the sky produced by VIRCAM with its 16 non-contiguous chips (see Fig 2-2 of [AD1]). The name is from the similarity to the prints made by the padded paw of an animal (the terminology was more appropriate to 4-chip cameras).

Tile A filled area of sky fully sampled (filling in the gaps in a pawprint) by combining multiple pawprints. Because of the detector spacing the minimum number of pointed observations (with fixed offsets) required for reasonably uniform coverage is 6, which would expose each piece of sky, away from the edges of the tile, to at least 2 camera pixels.

2 The UK User Requirements

In what follows we list all of the requirements specified in the UK User Requirements Document as a series of tables divided into categories as in the original document [AD3]. The index in the first column refers to the chapter and requirement number. The text in italics is a précis of the requirement (the reader is referred to the original paper for a more complete description). The regular text is the CASU response to the requirement. Only requirements that are relevant to CASU are commented upon.

2.1 General requirements

5.1	*All VISTA science data to be processed by VDFS*	Covered. Discussed in section 4.
5.2	*VDFS will respect all proprietary periods on data*	WFAU
5.3	*Archive output to be VO compliant*	WFAU/CASU, though there is a worry about agreeing to standards about which there is no firm community consensus.
5.4	*Pipeline will keep up with the data rate*	Covered if we buy enough hardware. See section 4
5.5a	*Reduction history written to FITS header*	Covered. See [AD2], chapter 5.
5.5b	*Reduction history sufficiently documented to reproduce reduced results.*	The comment is covered by 5.5a, *i.e.* there will be a traceable series of processing steps recorded in the headers that provide enough information to do a comparison given the relevant technical description
5.6	*Pipeline can cope with dead detectors*	Covered. This is fundamental to the design. See [AD2], chapter 5.
5.7	*Survey processing progress tools to be made available*	We have developed a system for WFCAM to do this and can apply similar methods for VISTA.
5.8	*Provision of summary statistics for data quality assessment*	Covered. See [AD2], chapters 10-11.
5.9-17	*Archive query requirements*	WFAU

2.2 Astrometric Requirements

6.1	*Absolute astrometric accuracy ≤ 0.3 arcsec*	Covered. WFCAM is regularly achieving < 100 mas accuracy and there is no obvious reason why VIRCAM should not do the same. See [AD1], section 3.2 and [AD2], sections 2.9 and 2.13.2.
6.2	*Differential astrometric accuracy ≤ 0.1 arcsec*	
6.3	*Differential astrometric accuracy ≤ 0.03 arcsec for a detector*	Both are covered provided sources have high enough signal:noise and provided intra-pixel effects, colour-dependent and other atmospheric effects are not an issue.
6.4	*Differential astrometric accuracy ≤ 0.01 arcsec for a detector*	

2.3 Photometric Requirements

7.1	*Absolute photometric accuracy ≤ 0.02 mag in J,H,K*	Both are covered, but the 1% goal may be impossible in practice. This will depend on the quality of the input data. WFCAM is regularly achieving the 2% requirement. See [AD1], section 3.3.
7.2	*Absolute photometric accuracy ≤ 0.01 mag in J,H,K*	
7.3	*Absolute photometric accuracy ≤ 0.03 mag in Y and z'*	Covered, though it is difficult to verify completely to this level due to a lack of external standards.
7.4	*Pipeline should handle defocused bright objects*	This goal is hard, since if you defocus too much it is difficult to measure the aperture corrections accurately and automatically. We also question whether this will ever actually be necessary since 2MASS will be used to calibrate each field. How the request to defocus the telescope will work its way into the input OB is also an interesting question.

2.4 Tiling, Stacking, Microstepping

8.1	*Combine images into pawprints with clipping*	Covered. See [AD2], section 7.11
8.2	*Image combination should also output a map of the clipped RMS for each pixel*	This would require major modifications of existing software, and we have severe doubts about its usefulness. Surely it would be better to do variable objects from either catalogues or difference imaging, since, at the pixel level, all objects vary due to seeing differences

		and inherent pixelation.
8.3	*Image stacking should be done with choice of several interpolation schemes.*	From a pipeline point of view the same stacking for all is preferred. Otherwise, as we have noted previously, the information about which stacking algorithm to use has to be propagated all the way through the observing system (see section 3.3). The choice of stacking method would have negligible impact on derived QC parameters. Drizzle, SWARP, MONTAGE *etc* are available as external packages so cost-wise it would be better to run different stacking/tiling choices on the archived products. There also appears to be confusion in the document about whether interpolation is required when forming the pawprints (jittering) as well as tiling. The jitter offsets should be sufficiently small so that resampling may not be necessary when forming the pawprints. If possible, we would prefer to use a nearest neighbour approach for the jitter stack and reserve any interpolation for the later tiling step.
8.4	*It should be possible to interpolate microstepped observations onto a smaller grid*	This would require modifications of existing software, although these are already planned and needed elsewhere.
8.5	*Should be able to handle 2×2 and 2×1 microstepping*	The 2×2 case is covered ([AD2], section 6.12). A significant change would be required to handle 2×1 interlacing, and the 45^{o} implicit rotation requires more thought as it will have interesting consequences.
8.6a,b	*Tiling should allow option for trimming or retaining of overlap regions.*	If a choice is really needed, this generally moves tiling to the archive end. Providing such an option at the pipeline end requires that the choice be propagated through the OBs into the header. (See section 3.3.) But why would you want to 'trim', since you are throwing information away? Which contributing pixels would you choose? Since the associated confidence maps convey all this information anyway what's the point? (It's effectively no different to taking the normal stacked

		pawprints and adding noise to higher confidence pixels). Another possibility is to modify the catalogue generation software to deal with this (in fact it already does). Option b is covered.

2.5 Variable Objects

9.1	*Flag probable variables when co-adding multiple epochs*	WFAU. Catalogues or difference imaging could be used for this - see comment to 8.2.
9.2	*Reject outliers from queries where significant time lag exists between OBs*	WFAU
9.3-4	*Suggestions for time series queries to search for variables*	WFAU: Time series analysis is best done as a list-driven photometry application from a master catalogue at the archive end.

2.6 Object Catalogues

10.1	*Catalogues include all photometric and astrometric parameters*	Covered. See [AD2], section 5.12.
10.2	*Catalogue content contains WFCAM parameters and possibly more*	
10.3	*Provision of catalogues for single and multiple band observations*	WFAU
10.4	*Flag solar system objects*	WFAU
10.5	*Flag known asteroids*	WFAU: Not all asteroids are catalogued and the surveys will turn up many more than are currently known about, which implies this type of detection/flagging is best left until later catalogue matching stages
10.6	*10σ point source completeness is 99.5%*	This is probably covered, though we would need to assess it with real VISTA data. It will require a good jittering strategy.
10.7	*8σ point source completeness is 90%*	This probably requires a more sophisticated detection filter as it is specialist LSB galaxy detection. The feasibility would need assessing with real VISTA data. [10-sigma at 4.5x is probably

		currently met]
10.8-9	*10σ point source completeness and reliability is 99.9% for well sampled objects that appear on good stacked images*	This is notwithstanding the currently unknown foibles of the detectors.
10.10	*Provide a completeness estimate for each tile*	This is in planned design enhancements for WFCAM.
10.11	*Provide a full Monte-Carlo completeness estimate*	WFAU: This may be tricky if the system does not have reasonably uniform sensitivity (WFCAM detectors have QE variations of a factor of 2). Assessing this will only be possible once we have real data.
10.12	*Basic variability using friends to friends algorithm*	WFAU. Is this the same as 9.3?
10.13	*List driven photometry*	WFAU. Is this the same as 9.4?

3 Pipeline Processing

3.1 Overview Of Standard Processing Steps

The processing steps that need to be applied to VIRCAM data have been discussed at length in [AD1] and their implementation in [AD2]. Below is a very brief description of each reduction step (in the approximate order of occurrence) as it applies to a set of target observations.

- **crosstalk correction:** Crosstalk images are removed from the input image (this may only be possible after flat fielding).
- **dark correction:** Dark current is removed from each image by subtracting an appropriate mean dark frame.
- **linearity correction:** Input images are linearised by applying a known polynomial expansion for the linearity curve and timing information for each frame.
- **flat fielding:** Input images are corrected for small and large scale throughput variations by dividing each image by an appropriate mean twilight flat field image.
- **background/sky correction:** Input images are corrected for spatially variable additive background using one of several methods described in the next section.
- **defringing:** Fringes are removed from the input images by scaling and subtracting master fringe frames. (There is the distinct possibility that this will not be necessary.)
- **decurtaining:** Low level stripes caused by electrical interference are removed from the background.
- **persistence:** Persistent images are subtracted using a decay model and a known time constant.
- **interleaving:** Microstepped frames are interleaved onto a finer grid using known offsets
- **jittering:** Objects on frames that have been jittered are located and cross-referenced to calculate jitter offsets. The offsets are used to shift and combine the input images into a single stacked image.
- **catalogue generation:** All astronomical objects on the output stacked image are catalogued and classified.
- **astrometric calibration:** The objects in the catalogue are matched to astrometric standard stars (2MASS). A plate solution is found defining the image world coordinate system (WCS).
- **photometric calibration:** The objects in the catalogue are matched to photometric standard stars (2MASS). A photometric zeropoint is calculated.

3.2 Extra Processing

Because the summit and Garching pipelines only work causally with observation time, the best choice of calibration information which is relevant to a particular observation may not be available in time to reduce a given OB. In this case master calibration frames/tables will be used, which means that there will be some reduction steps that are not done optimally in these pipelines. The Cambridge pipeline will allow an entire night (or more) to be treated as a single entity and thus allow OBs to be reduced in a different order from when they were taken, ensuring that the most up-to-date calibration information is used. In this section we explain the major benefits of this ability to juggle the reduction order.

3.2.1 Background Correction

Background correction is one of the most difficult effects to compensate for in infrared imaging. Even small background variations may seriously affect the sky correction for extended objects. We define 'background' here as a combination of additive effects that are both spatially and temporally variable. This is distinct from any spatially variable background light originating from an astronomical source (say the halo of a bright galaxy). There can be many sources of background emission – some of these are:

- Spatially varying night sky glow which originates in the atmosphere. This can often been seen on spatial scales much smaller than the projected area of a detector.
- Scattered light in the camera
- Fringes caused by interferences of night sky emission lines with the top layer of the detector material
- Thermal emission originating from dust collecting on optical surfaces (this is especially prevalent in the K band).
- Stripes caused by electronic interference.

Because these effects all vary with time, the only source of information we really have for reducing their impact on an OB is the data themselves. The spatial variation of the background means that it generally has to be modelled in two dimensions and this can be very difficult on frames that contain extended objects or in very crowded fields. In what follows we describe how we intend to improve the background correction for each of these cases.

3.2.1.1 Sky Glow, Thermal Emission and Scattered Light

Most observations will contain large areas of 'blank sky', and most will also be jittered in an effort to remove bad pixels from the final map. These two facts are key to the methods that are used to remove sky glow from images. As described in [AD1] and [AD2] the summit and Garching pipelines will do a two dimensional estimate of the sky by combining with rejection all of the frames from an OB. Because the images are normally jittered and there are lots of patches of blank sky, it is almost always the case that the astronomical objects will be rejected from the stack and we will be left with an image of the sky glow, thermal emission and any scattered light. This will be normalised to a zero median and then removed from the frames themselves.

In the situation where an OB is flagged as having extended objects, this method can't be used, as the object sizes will probably exceed the characteristic jitter offset and features such as these will not necessary be rejected completely during the stacking phase. The same is true for crowded fields. In this situation it is common practice to observe an 'offset sky' region. That is a region which is unaffected by large extended objects and is not in a crowded region, and yet is spatially close enough to the programme region so that any background variations arising from the atmosphere or scattered light will be roughly the same. The offset sky observations will be taken either just before or just after the observations of the programme region, so that any differences between the sky glow measured from the offset region compared to that which exists for the programme region will be largely unaffected by the temporal variations that are known to occur in the atmosphere. Observing the offset sky region close in time to the programme region is also essential for removing thermal emission as even small changes in the ambient temperature can cause a variation in the thermal emission relative to the sky glow.

The summit and Garching pipelines will not be able to generate a background correction in the situation where an offset sky is needed. This is because there is no way of knowing a priori which offset sky region is attached to which target field, and it may also be that when it comes to reducing the former, the latter has not been observed yet. Doing a standard background correction with stacked programme images would probably make the situation worse, as mentioned above. Although leaving the images uncorrected for the background is not ideal, it is still a sufficient level of reduction for calibration and quality control, which is the purpose of the summit and Garching pipelines. Because we are able to shuffle the order in which we reduce OBs in the Cambridge pipeline, this problem goes away. OBs that are flagged as 'offset sky' regions are reduced as any other region would be, except that they will output a background frame as a data product. These will be reduced before any of the relevant target OBs are done, ensuring that an appropriate offset sky estimate will be available for OBs that need them.

3.2.1.2 Fringes

At the time of writing, we have no way of knowing what effect fringing will have on data from VIRCAM. Some infrared detectors are badly affected by it (UFTI) and some are not affected at all (WFCAM). It appears to depend upon the final f-ratio of the optical system, the properties of the top layer of material in the detectors and the presence or not of atmospheric emission lines in a particular waveband. It is also true that a background sky correction may remove any fringing that is present if the sky estimate is sufficiently local both spatially and temporally. Document [AD2] gives a thorough explanation of how fringes can be removed from images. The basic method is to fit the fringe pattern from a library fringe frame to that of an observation frame by iteratively minimising the median absolute deviation of the difference of the two images. This should work in principle so long as the fringe pattern is stable with time. However, experience shows that this is not the case. The flux of the emission lines

that lead to the fringe patterns can vary in a complex temporal manner which means the relative intensity of parts of the fringe pattern will also alter with time.

The way to get around this problem is to use data from the night in question to form mean fringe frames rather than to rely on a library frame which may be days or even weeks old. As this can only be done once the whole night of data has been at least partially reduced, this method of fringe correction will only be possible in the UK pipeline.

3.2.1.3 Stripes

Laboratory VIRCAM frames suffer from low level striping perpendicular to the detector readout axis. This is an electronic effect that is different for every detector and every exposure, meaning that there is no way of having a standard calibration 'stripe image'. Modelling this out involves calculating the median background value for each row in an image, ignoring bad and object pixels. The one-dimensional stripe profile is block replicated to a two-dimensional image and subtracted from the input image. At present we have no way of knowing the level of striping that will appear on the observed images when on the telescope due to the difference in electronic environment between RAL and Paranal. It may be that, with appropriate tuning of the system, these are at a level where they can safely be ignored.

In the event that we have to remove this effect, there is a facility in the summit and Garching pipelines to do it. Obviously the presence of large extended objects and crowded fields will make it impossible to estimate some of the medians properly and in that case doing this correction will actually degrade the quality of the background image. In this case, as with the sky glow subtraction, the Garching and summit pipelines will not attempt this. In the Cambridge pipeline we will be able to flag object pixels when forming the medians and hence get a better background estimate.

3.2.2 Tiling/Resampling

The layout of the detectors in the focal plane of VIRCAM and the desire for nearly uniform observational coverage of a contiguous area of sky has lead to an offsetting strategy of six pawprints, as discussed in [AD3]. Done correctly, a full tile observation will ensure nearly uniform coverage and each area of sky will be observed at least twice. A result of this strategy is that in order to extract objects to the depth specified by a given survey, it is necessary to stack the resulting pawprints into a filled tile image. (A single pawprint will only have half the depth of the full tile.) The distortion of the focal plane by the camera optics also requires that in order to create such a tile, input pawprints will have to be resampled onto a new grid.

The Cambridge pipeline will stack output pawprints into a final tile image before extracting a final catalogue. This is step is in addition to those outlined in section 3.1 of this document. A finite number of resampling algorithms will be provided by the pipeline and a choice can be specified by the observer when creating the OBs for his survey. Tiling will also involve re-gridding some pixels on the outer parts of the focal

plane as the optical distortions in the camera cause the projected area of sky per pixel to vary across the field. It is envisaged that the tiles will be regridded onto a tangent plane projection (see [RD 3] for more information) which provides a nearly uniform projected area for all pixels.

3.3 The RECIPE Keyword

Many of the extra processing stages described in this document rely on information from the observers about the nature of the observations. Information regarding the type of background correction to be done or the interpolation routine to be used in stacking needs to be specified in order for the pipeline to run without manual intervention. This will be done using the RECIPE keyword, which has been specified for the primary header unit of each of the VIRCAM FITS files. The header value for RECIPE will consist of a comma separated list of words that describe the extra steps that are to be taken when reducing a specific OB. A value of DEFAULT will trigger the VIRCAM standard reduction. A full list of allowed values will be published in due course.

4 Cambridge Pipeline Infrastructure

4.1 Data Transport

Transmission of around a third to a half terabyte a day from Paranal to Europe is currently not possible using networking. Data is written to hard discs using a Linux operating system and the ext3 file system. The discs are physically located in a Chenbro chassis and are permanently mated to the corresponding mounting bracket. Currently, a capacity of 250GB is standard. Arrangements are made to ship these discs periodically to Garching, where they are inserted into an identical Chenbro chassis and the data read off.

Subsequently, because the network connection to ESO headquarters is also insufficient to support bulk data transfers, discs will be shipped to Cambridge for insertion into another similar chassis and read again for use in the UK pipeline. It is expected that there will be a three-week cycle before a particular disc is returned for re-use.

With regard to transporting data from Cambridge to Edinburgh, the situation is much simpler. With gigabit connectivity either over JANET or the reserved "UKLight" connection, reduced science products will be flagged as ready-to-transfer and copied up to the WFAU archive automatically, using high-performance networking protocols (see [AD8]).

4.2 Data Ingestion

Data ingestion is one of the most manually-intensive tasks in the pipeline, taking non-trivial operator effort. A substantial percentage of this time will undoubtedly be spent in recovering lost data. The routine cycle of data ingestion will consist of the following steps and is illustrated in Figure 4-1:

- **Copy ESO Discs**: The data ingestion portal will consist of a relatively modest Linux system using a Chenbro chassis into which we will plug the ESO delivered discs. Data will be copied over to raw storage via a gigabit network using the NFS protocol. A standalone ingestion system is desirable because the constant physical insertion/removal of discs will inevitably result in frequent system reboots.
- **Verify Raw Data**: Following copying, the new raw data collection will be verified by several means, including checking the FITS MD5 checksum, running "fitsverify" and cataloguing the files to compare with the expected list of observations. Verification failure will trigger a recovery process which would vary depending on the scale of the data dropout; a single file might be transferred over the network, a whole disc failure would require a resend from Garching or Paranal as appropriate.

- **Backup Raw Data**: The opportunity exists at this stage to backup the raw data onto archival media, for example LTO-n tapes in a tape robot. This is considered necessary and will be done at regular intervals.
- **Recycle Discs**: Shortly after the integrity of the new batch of raw data are affirmed, the transport discs are replaced into their shipping container and sent back to Garching.

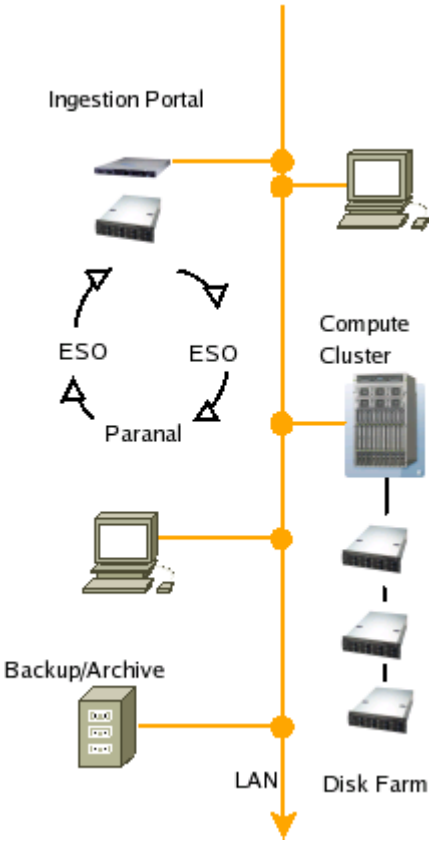

Figure 4-1 An illustration of the flow of data from Paranal to Cambridge

4.3 Triage

By this stage the data are safely on disc as an exact copy of that which left Paranal. However there will certainly be a fraction which is unusable for example underexposed or saturated twilight flats which will need pruning out. There are several ways this can be done, for example:

- **Indexing**: The first step after ingestion is to create an index of each night of data. Scripts will be used to group data into OBs and flag all those that are deemed unusable (e.g. data taken during the day for engineering tests).
- **Bad Frames by QC**: The QC pipeline would generate parameters indicating problems, and if these are routinely available, it may be possible to pre-identify bad frames. At any rate, the UK pipeline will be generating exactly the same set of QC parameters and so after reprocessing, frames producing

outlying QC parameters will be examined and, if necessary, flagged as unusable.

- **Bad Frames by Inspection**: In particular, it is essential that frames contributing to calibration data, i.e. darks and flats are high quality. Unlike the near-time QC pipeline, the UK pipeline has the advantage of being able to use data from both before and after a particular science observation. Great care is necessary in building the calibration products and skilled human intervention is a necessary part of the process.

4.4 Software

Data reduction pipelines have three basic components.

- A data access component. This is what allows access to input/output FITS images/tables and their headers.
- A core of data reduction modules. These are what do the actual work on the data.
- An intelligent 'glue' that is used to bolt the reduction modules together into a pipeline.

These three components are not necessarily separate in that there is often overlap between the functions provided by them (e.g. there is often a need to access information from a FITS file within the pipeline glue).

CASU has written a great deal of production pipeline software using the package **cfitsio** for its data access needs. This is as close as we can get to an 'industry standard' data access package for astronomy. It is used directly in the data reduction modules to do the entire file I/O for FITS images and tables. The CASU infrastructure then wraps the data reduction modules using the **XS** facility in **Perl** – this makes them callable from within Perl (which also has access to cfitsio through an XS wrapper). Perl routines are then used as the glue to organise the data and present it to the reduction modules.

ESO, on the other hand, have their own infrastructure called **CPL** (see [RD 4] and [RD 5]). This consists of libraries of routines for image and table i/o as well as some basic data manipulation algorithms. There are also higher level facilities within CPL which allow you to create plugins for ESO user interfaces. At the heart of CPL is **qfits,** which is an ESO specific FITS package. CPL is written in such a way that the user is not allowed to call qfits directly, but rather must call an appropriate CPL routine to access FITS data. This was done to encapsulate the data file i/o away from the user so that CPL could, in future, seamlessly handle data of several different formats.

A decision was reached in conjunction with ESO that the summit and Garching pipelines should be written in the CPL environment. This meant that the CASU modules that had been developed over many years had to be recast to use the CPL infrastructure. (However, they are still functionally identical to the CASU versions).

Although the CASU infrastructure is far more flexible than CPL, to use it in the UK pipeline would mean having to maintain two separate sets of data reduction modules. This would clearly not be a cost effective or efficient approach and it has thus been decided that the bulk of the Cambridge pipeline will reuse the modules delivered for the ESO pipelines.

The high level Data Organiser used at ESO to sort data files and present them to the pipeline will not however be used in Cambridge. It is too restrictive and changes to it would be difficult or impossible. Instead a set of Perl scripts, using the CASU infrastructure will be written to create the reduction menu to present to the pipeline. This will allow for rapid development of the organisational procedures and will not affect the status of the CPL based reduction modules.

4.5 Hardware

The requirements for processing power can be based on experience with WFCAM and from benchmarking results from early versions of the VIRCAM reduction recipes. Storage capacities are based on the likely data-acquisition curve for VISTA. The estimate of 210 Gbytes/night contained in [AD3] suggests an annual data rate of about 75 Tbytes/year. Rice tile compression will lower this to about 25 Tbytes of physical storage for raw data. The UK User Requirements [AD3], requirement 5.4, also states that the pipeline should be able to operate at 1 Tb/day for 10 days, and sufficient headroom must be available to allow scheduled and unscheduled downtime (hardware and software upgrades, system failure recovery). Given the well-known projection of future price/performance in IT, even at this stage the exact specification of equipment cannot, and should not, be made. Final choices of actual hardware will be based on a total systems approach, taking into account not only simple processing per pound, but physical footprint (real estate), power requirements (including cooling), reliability and serviceability.

4.5.1 Processing Power

Enough processing power must be available to meet the peak sustained data rate of 1Tb/day stated above. Operational experience with over a year of WFCAM reductions have shown that and 8-cpu (4 3GHz twin processor) array can handle both regular data deliveries and re-processing with improved techniques. The VISTA data rate is expected to be (only) twice that of WFCAM on average and so it is expected that a 16-cpu system will cope with the demands.

The ESO DFS group have benchmarked the QC pipeline [RD 6] and derived a requirement of 16 CPUs to cope with near-time data analysis. The base model tested is a 2.4GHz Opteron with 4Gb memory. While that system must be able to run at peak data rates, and the UK one only at mean data rates, the UK pipeline will be doing extra processing compared with the QC pipeline. If these two effects balance out, then 16 CPUs of this specification can be taken as a good independent recommendation. Added to the baseline requirements will be redundancy machines, disk servers, and ingestion and export portals.

In practice, an over-specified system will be installed to allow for hardware problems, reprocessing and new software testing.

4.5.2 Disc Storage

Enough disc storage must be available to hold the likely mean data acquisition stated above. A reasonable purchasing model might be to acquire enough discs to hold a quarter's worth of raw and reduced data at once. Given a factor of 2 expansion in reduced frames we would need 3 times the raw data volume. Luckily, though, this will compress by a factor of 3 and so the first tranche of operational disc storage (covering the first semester or so) would be around 30Tb. At the time of writing this might be in the form of 3×10Tb SATA RAID6 fibre channel arrays costing £1k/Tb but as usual the cost is expected to reduce in the long run.

4.5.3 Off-line archive

Having a duplicate copy of the reduced data products at WFAU is probably the most effective way to have a much desired backup. This is tantamount to making a disc copy and storing it off site.

4.6 Human Resources

The pipeline will operate nearly continuously and although it is massively automated, at any one time at least one person will need to monitor progress and perform interactive quality control. The data ingestion cycle takes about 3 weeks, within which in a given week there is about a day of disc handling. Data-transfer disc problem-shooting might take a day a month. Hardware commissioning and maintenance will be required and so will operating-system maintenance and upgrades. Upgrading, maintenance and testing of new releases of VDFS software will also take significant time and effort.

Finally of course allowance must be made for management and administrative overheads. Operations staff members are expected to participate in relevant external meetings and conferences, and exploit training and development opportunities.

At CASU VDFS development is currently funded at the 3FTE level until the end of September 2007. Because of the delay in VISTA commissioning we are negotiating with PPARC for a 6 month extension of the development effort, since some of it is contingent on early commissioning results and more detailed understanding of the properties of the VIRCAM detectors.

Further operational development and the running costs of the VISTA pipeline in Cambridge will be sought as part of the CASU rolling grant renewal, with a start date of 1 April 2008. From previous experience we anticipate VDFS operating requirements to be at the 3 FTE level.

5 Risk Register

A risk register for the VDFS is given in the table below. Given the mature state of the VDFS design, most of the risk areas are well understood. The 'likelihood' and 'impact' of the risk are both rated on a scale of 1 to 4 (the latter being the worst case). The total 'exposure' of the risk is rated as the product of these two.

ID	Owner	Monitor	Risk Summary	Risk Details	Area	Mitigation Strategy	Likelihood	Impact	Exposure
1	VPO	JRL	Detector	Unforeseen detector peculiarities	Camera	Liaise with VIRCAM team	3	3	9
2	IOA	MJI	New Hire	Recruitment delays	Personnel	Temporary reassignment	3	1	3
3	IOA	PSB	Key Staff	Loss to other lucrative or secure jobs.	Personnel	Replace + make sure good documentation is in place	2	4	8
4	IOA	JRL	Software	Problems in development of new algorithms	Software	Be smarter	2	2	4
5	IOA	JRL	Reprocessing	Bug fixes and improvements to algorithms	Software	Allowed for with accurate benchmarking	4	3	12
6	IOA	PSB	Hardware	Setup and delivery delays	Hardware	Commodity products	2	2	4
7	IOA	PSB	Disk failure	Loss of raw data due to disk crash/failure	Hardware	Raid systems and offline backup will be used	4	1	4
8	IOA	PSB	Disaster	e.g. Fire wipes out APM building	Hardware	Data and software backup kept external to APM	1	4	4
9	IOA	PSB	Data volume	Data transfer problems	Media Network	Purchase more capacity	3	4	12

6 Schedule

We will discuss this at the review.

_o☺o_