

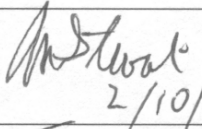
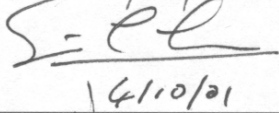
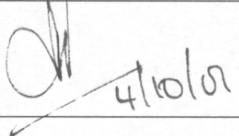
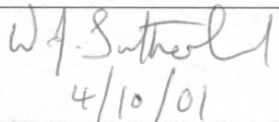
Tel: +44 (0)131 668 8411
Fax: +44 (0)131 668 8412
Email: vista@roe.ac.uk,
WWW: <http://www.roe.ac.uk/atc/vista>

Document Title: VISTA Data Storage and Transport

Document Number: VIS-TRE-ATC-00150-0005

Issue: 2

Date: 2 October 2001

Document Prepared By:	J M Stewart Software Engineer	Signature and Date:	 2/10/01
Document Approved By:	S Craig Project Engineer	Signature and Date:	 4/10/01
Document Released By:	A McPherson Project Manager	Signature and Date:	 4/10/01
Document Reviewed By:	W Sutherland Project Scientist	Signature and Date:	 4/10/01

The information contained in this document is strictly confidential and is intended for the addressee only. The unauthorised use, disclosure, copying, alteration or distribution of this document is strictly prohibited and may be unlawful.

Change Record

Issue	Date	Section(s) Affected	Description of Change/Change Request Reference/Remarks
1	24 June 2001		Original presented for Close Out Review as VIS-SPE-00150-0005
2	26 Sept. 2001	3.1, 3.3.2	Data rates updated. Topology diagrams clarified.

Table of Contents

1	INTRODUCTION	3
1.1	PURPOSE.....	3
1.2	SCOPE.....	4
1.3	APPLICABLE DOCUMENT	4
1.4	REFERENCE DOCUMENTS	4
1.5	ABBREVIATIONS AND ACRONYMS	5
2	CONSTRAINTS	5
3	ON-LINE STORAGE	5
3.1	REQUIREMENTS	5
3.2	TRENDS	6
3.3	SOLUTIONS	6
3.3.1	<i>Topology</i>	7
4	DATA TRANSPORT	9
4.1	REQUIREMENTS	9
4.2	SOLUTIONS	9
4.2.1	<i>DVD</i>	9
4.2.2	<i>Magnetic Disk</i>	10
4.2.3	<i>LTO Tape</i>	10
5	NEAR-LINE STORAGE	10
5.1	REQUIREMENTS	10
5.2	SOLUTIONS	10
6	RECEPTION OF DATA IN UK	11
6.1	REQUIREMENTS	11
6.2	SOLUTIONS	11
7	CONCLUSIONS.....	11

1 Introduction

1.1 Purpose

This document examines how VISTA's requirements for data storage and data transport can be met. VISTA's demands are large in both areas and it is necessary to assess the potential technical and economic challenges. Since VISTA is scheduled not to start producing scientific data for some years, an assessment is made of the future performance and costs of the available technology.

1.2 Scope

Data storage is considered at Paranal i.e. excluding any requirements or goals that may exist for Garching or within the UK.

Data transport is considered from Paranal to Garching. It is assumed the same medium will be used between Garching the UK.

The receipt of data at the UK data centre is a VISTA Project responsibility and is considered. However subsequent data handling at this site is not considered.

1.3 Applicable Document

- AD01 *VISTA Science Requirements*, VIS-SPE-VSC-00000-0001, V2.0, 26 October 2000.
- AD02 *VISTA Technical Specification*, VIS-SPE-ATC-00000-0003, Issue 2, 26 September 2001.

1.4 Reference Documents

- RD01 *VISTA Instrument Software Requirements*, VIS-SPE-ATC-00150-0003, Issue 2, September 2001.
- RD02 *VISTA Software Architectural Design*, VIS-SPE-ATC-00150-0001, Issue 2, September 2001.
- RD03 *VISTA Computer Hardware Architectural Design*, VIS-SPE-ATC-00150-0002, Issue 2, September 2001.
- RD04 *Final Layout of VLT Control LANs*, VLT-SPE-ESO-17120-1355, V1.2, 12 Jan. 1999.
- RD05 *The Decline of Magnetic Disk Storage Cost over the next 25 Years*, Archive Builders, <http://www.archivebuilders.com/aba/004.html>.
- RD06 *Fluorescent Multilayer Disks Whitepaper*, Constellation 3D Inc., 7 June 2000, <http://www.c-3d.net/downloads/whitepaper.pdf>.

RD07 *HP Ultrium Technology White Paper*, Noveember 2000, HP Inc.,
<http://www.h.com/storage/> .

1.5 Abbreviations and Acronyms

ESO	European Southern Observatory
LTO	Linear Tape Open
NAS	Network Attached Storage
SAN	Storage Area Network
VLT	Very Large Telescope
VST	VLT Survey Telescope

2 Constraints

VISTA will be operated and maintained by ESO staff and so all systems installed at the Paranal site are required to comply with ESO's requirements and constraints. ESO perform data storage and transport functions for the VLT's instruments and, when commissioned, also for VISTA and the VST. VISTA will therefore comply with ESO standards as they exist in 2004 and will, where appropriate, provide input as these standards evolve.

Because of the need to comply with ESO standards and the fact that this area of technology is evolving rapidly, means that it is not feasible to prescribe in 2001 exactly what should be operational in 2006. However options and cost estimates will be described. As will be demonstrated, there are several technical solutions even today.

3 On-Line Storage

3.1 Requirements

The basic requirement is on-line storage for 2 typical nights or, if greater, one worst case night AD01. This translates to 0.8 TB i.e. 2 typical nights with the IR Camera AD02.

The continuous worst case data rate that must be handled is 54 MB/s AD02.

3.2 Trends

Magnetic disk storage continues to improve rapidly in terms of capacity and performance. One analysis quotes the cost/GB decreasing by 37.5% per annum RD05, analogous to Moore's Law. Therefore by mid-2005, when the final operational hardware might be purchased, the capital cost would only be 20% of today's cost.

3.3 Solutions

3.3.1 Technology

On-line storage implies magnetic disks. Further considerations include fault tolerance and whether the data should be hosted by a single system or made available to several.

A RAID array is a common means of providing such storage. Compared to single disks, RAID arrays offer greater performance, due to parallel data flows, and fault tolerance, since the data can be spread across several drives with parity or data replication

Although 1 TB of magnetic disk storage may seem large, by the standards of 2001 it is already fairly routine. This is because the data explosion is a phenomenon not confined to astronomy, but one that affects many areas of science, engineering and business, so fuelling technical development and cost reductions.

There are many RAID products on the market, differentiated, e.g., by performance, expandability and brand name. Some examples of what is currently on the market are given in Table 1, with prices from public price lists.

Table 1 Some disk systems currently on the market.

Description	Capacity	Vendor	Price (mid-2001)
181 GB SCSI disk	180 GB	Eclipse	£1.4k
Barracuda 180GB	180 GB	Seagate	\$2k
FC60 RAID system with 10 SC10 enclosures and 38x 36 GB disks, Fibre Channel	1.2 TB	HP	\$155k
RS15-R1200 RAID array, NAS	1.03 TB	Raidzone	\$21k

ESO have chosen to use HP workstations at Paranal and so the reasons why the HP RAID system appears so expensive should be commented upon. The HP system is expandable to 100 disks in a rack and several racks can be integrated into a single system. The HP system has two Fibre Channel interfaces. Brand name probably has an effect. However the cost is largely dominated by the costs of the disks themselves, which HP price at \$2k for 36 GB (\$55/GB), as opposed to Seagate's \$2k for 180 GB (\$11/GB). When tendering for a real

system, one would likely receive a quote based on the current market, rather than the price in a catalogue.

The worst case continuous data rate of 54 MB/s, like the data volume, is well within the capabilities of today's technology and will be commonplace by 2005. RAID systems are often supplied with Fibre Channel interfaces to the host(s). Fibre Channel supports a data rate of 100 MB/s and the RAID arrays are designed to be able to handle this using parallel data paths and data buffers. (Throughput can be increased by using multiple Fibre Channel interfaces, but this will not be necessary for VISTA.)

A suitable system for VISTA could include:

- 1 TB
- 2 Fibre Channel interfaces
- no single point of failure
- hot swappable disks and power supplies
- RAID 5 fault tolerance

From a mainstream supplier offering support, e.g. spares and maintenance software, the cost in 2005 might be of order \$40k.

3.3.2 Topology

The network topology currently deployed at Paranal RD04 implies storage attached to workstations (note that in ESO and HP terminology a workstation could also be a server). Data are transferred over the LAN, which is generally based on 155 Mbps ATM and switches. This topology is illustrated in Fig. 1. The IWS must accommodate 5 155 Mbps ATM interfaces to handle the data rates. The data themselves must be copied to each workstation that uses them. Although rather inefficient this topology will meet VISTA's requirements.

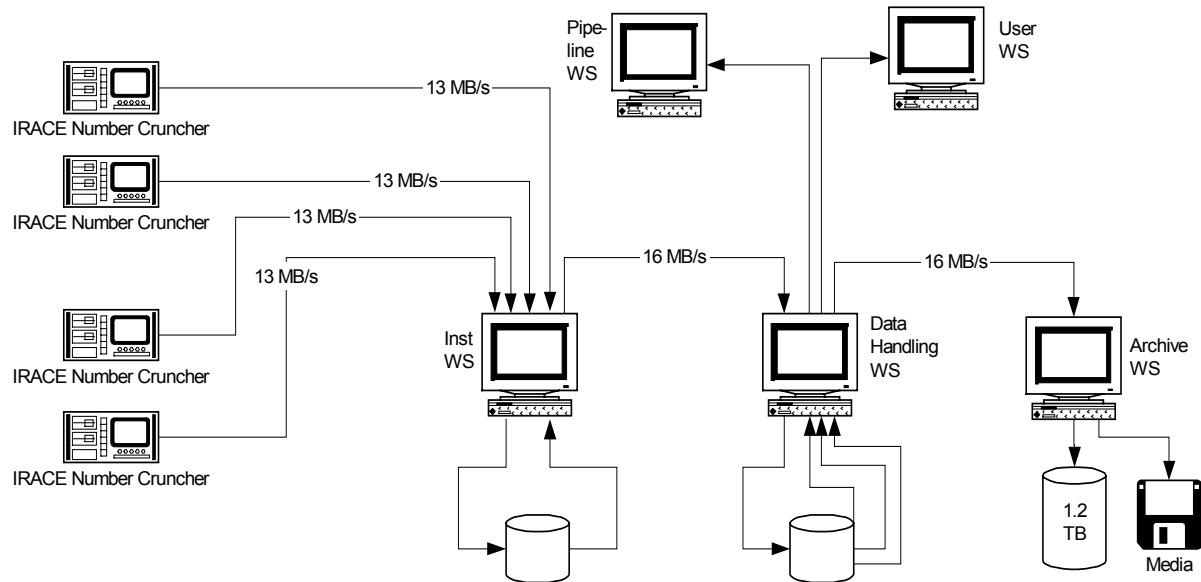


Figure 1 VISTA storage topology based on current VLT model using only 155 Mbps ATM.

The topology could be simplified in two ways. First the IWS could be connected to a single 622 Mbps ATM interface rather than 5 155 Mbps interfaces. Each IRACE Number Cruncher could be connected to the IWS via 155 Mbps ATM and an ATM switch. (A single Number Cruncher connected to 622 Mbps is another option.) Secondly Storage Area Network, SAN, technology could be used to allow the different workstations to access a common data store without physical copying of data. SAN commonly uses 100 MBytes/s Fibre Channel links using a protocol optimised for high volume data, rather than the multipurpose IP. (Network Attached Storage, NAS, is alternative to SAN, but currently does not provide adequate performance.) Use of both 622 Mbps ATM and SAN is shown in Fig. 2. The possibility of connecting the IRACE Number Crunchers to the SAN is not shown.

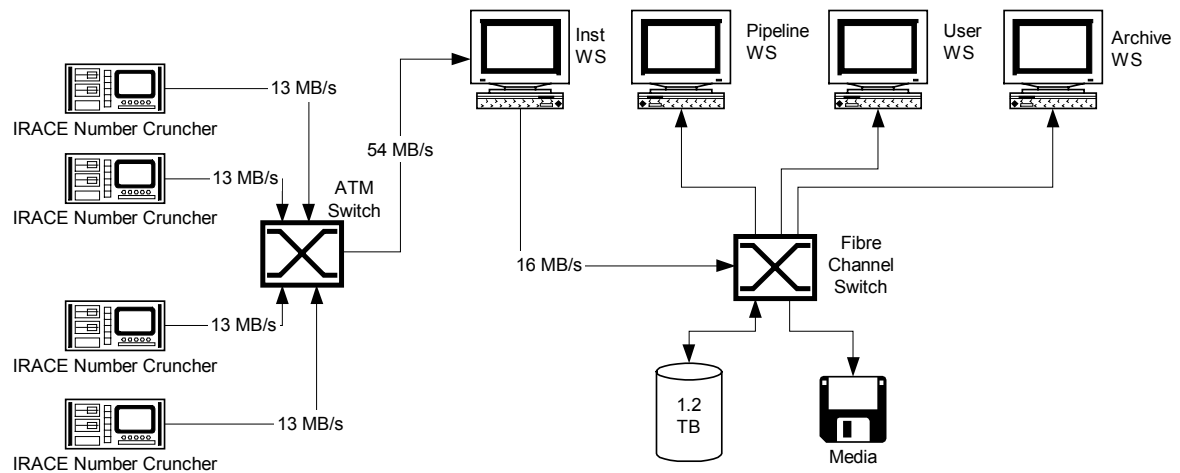


Figure 2 Alternative VISTA storage topology based using both 622 Mbps ATM (or Gigabit Ethernet) and Storage Area Network architecture.

4 Data Transport

4.1 Requirements

Electronic transmission of data from Paranal to Europe will not be economic at the start of VISTA operations and this may remain the case until the end of its life. Therefore data will be transported from Paranal to Garching using physical media. All raw data will be transported, e.g. 400 GB from a typical night with the IR Camera.

4.2 Solutions

4.2.1 DVD

Currently ESO use DVD for data transport. These media are robust and have had adequate capacity, 4.7 GB, for typical VLT instruments. However VISTA would require 85 media for a night's data, which although feasible would present operational problems. Cost of media for a night's data would be of order \$3k.

The data rates of a DVD RAM drive is approximately 1.3 MB/s. Assuming one can spend 24 hours writing data acquired in 12 hours of observing, one would need 6 drives. Such systems are available, e.g. the Plasmon DVD-RAM D480 JukeBox has up to 6 drives.

DVD capacity extends to 9.2 GB, but even this looks small by today's standards. Other optical technologies are currently being developed, e.g. Fluorescent Multilayer Disks are expected to offer 100 GB/disk RD06. However such systems are not yet on the market.

4.2.2 Magnetic Disk

ESO are currently exploring alternatives. Removable magnetic disks appear to be the option currently preferred. Using a 180 GB disk (see Table 1), one night's data could be stored on 5 disks at a media cost of \$10k. By 2004 one disk would have adequate capacity for a night's data at a cost of order \$2k.

4.2.3 LTO Tape

The tape format that seems most appropriate for this application is Ultrium/LTO, developed jointly by HP, IBM and Seagate RD07. Current systems can store 100 GB on a tape (105x102x21 mm) and there is a well defined development path that leads to 800 GB/tape over the next 5 years or so. Data transfer rates are currently 10 MB/s, rising to 80 MB/s.

An HP Ultrium 230 external drive is currently listed at \$5.5k and a tape cartridge costs \$90. Tape Autoloader and Library systems are also becoming available, e.g. the Overland Data LTO Ultrium Autoloader/Xpress stores 1.1 TB at a cost of \$8k. Plasmon's LTO Series under development will store 50 TB.

SuperDLT format is an alternative to LTO. This has the advantage of backwards compatibility (with DLT), but is not quite as fast as LTO and does not have such a clear future development strategy.

If one were to select a medium purely for temporary backup and data transport, LTO would be a front runner.

5 Near-Line Storage

5.1 Requirements

VISTA requires 30 nights' data to be available near-line RD01, which translates into 12 TB (IR Camera used for 30 typical nights each generating 400 GB).

5.2 Solutions

Near-line storage falls in the niche between on-line storage and data transport and could use either (or neither) technology. The cheapest way of meeting VISTA's requirement is probably to use a LTO Library/Autoloader (see 4.2.3) costing perhaps \$10k by 2004.

6 Reception of Data in UK

6.1 Requirements

The VISTA Project must provide a system to receive data at the UK Data Centre, but other costs will be found elsewhere, e.g. archival storage, processing and access.

6.2 Solutions

The same medium is likely to be used to receive data in the UK as to export it to Garching. Two copies could be made at Paranal or the data could be copied in Garching or the transport media could be forwarded to the UK after being read in Garching. the UK Data Centre will also need to receive data from other sources. It would be desirable, but probably not achievable especially over an extended timeframe, to use the same medium for all telescopes and data centres.

LTO would be a very most economic way to receive data in the UK. However the archiving centre may choose a medium other than tape for data archiving and it may be desirable to use the same medium for data transport.

7 Conclusions

- 1) There are no fundamental technical challenges is storing or transporting VISTA data.
- 2) There is no conflict with current ESO standards, but VISTA would benefit from either 622 Mbps or Gigabit Ethernet, in addition to or instead of the current VLT standard of 155 Mbps ATM.
- 3) RAID arrays are attractive for on-line storage.
- 4) Fibre Channel and SAN technology may provide benefits for VISTA's network and storage topology.
- 5) LTO tapes are attractive for data transport and near-line storage, but it may be desirable to use the same media as for archival storage.
- 6) Several solutions exist for near-line storage. LTO is probably the simplest and cheapest, but it may be appropriate to use the same medium as for archival storage in Garching and/or the UK. Using on-line storage throughout may become economic.
- 7) Costs will continue to fall rapidly over the development phase of VISTA. Purchases should be delayed as long as possible.

Doc Number:	VIS-SPE-ATC-00150-0005
Date:	2 Oct. 2001
Issue:	2
Page:	12 of 12
Author:	J M Stewart

oOo