VirtU - The Virtual Universe

comprising

the Theoretical Virtual Observatory & the Theory/Observations Interface

A collaborative e-science proposal between

Carlos Frenk	University of Durham
Ofer Lahav	University College London
Nicholas Walton	Cambridge University
Carlton Baugh	University of Durham
Greg Bryan	University of Oxford
Shaun Cole	University of Durham
Jean-Christophe Desplat	Edinburgh Parallel Computing Centre
Gavin Dalton	University of Oxford
George Efstathiou	Cambridge University
Nick Holliman	University of Durham
Adrian Jenkins	University of Durham
Andrew King	University of Leicester
Simon Morris	University of Durham
Jerry Ostriker	Cambridge University
John Peacock	University of Edinburgh
Frazer Pearce	University of Nottingham
Jim Pringle	Cambridge University
Chris Rudge	University of Leicester
Joe Silk	University of Oxford
Chris Simpson	University of Durham
Linda Smith	University College London
Iain Stewart	University of Durham
Tom Theuns	University of Durham
Peter Thomas	University of Sussex
Gavin Pringle	Edinburgh Parallel Computing Centre
Jie Xu	University of Durham
Sukyoung Yi	University of Oxford

Overseas collaborators:

Joerg Colberg	University of Pittsburgh
Hugh Couchman	McMaster University
August Evrard	University of Michigan
Guinevere Kauffmann	Max-Planck Institut fur Astrophysik
Volker Springel	Max-Planck Institut fur Astrophysik
Rob Thacker	McMaster University
Simon White	Max-Planck Institut fur Astrophysik

1 Executive Summary

1. We propose to construct a Virtual Universe (VirtU) consisting of the "Theoretical Virtual Observatory" (TVO) and associated diagnostic tools making up the "Theory/Observations Interface" (TOI). VirtU is a computing infrastructure to enable direct and rigorous comparisons of realistic simulations of cosmic structures, based on the best current theoretical understanding, with real data. VirtU will provide a set of tools to:

(i) Remotely access and manipulate a large, dynamic archive of simulated objects (galaxies, clusters, etc) using AstroGrid standards and procedures

(ii) Intercompare simulations carried out using different techniques

(iii) Intercompare simulated data held in the TVO with real data held in the Virtual Observatory

(iv) Enable parallel simulations to be executed remotely using grid technology

(v) Visualize multi-dimensional simulation output (including particle-based data) and observational data.

To achieve these goals, we propose to:

(i) Develop a standard data format for storing simulation output

(ii) Develop and grid-enable a set of powerful tools for the analysis of simulated and real datasets

(iii) Create a Virtual Telescope Simulator to aid in the planning of astronomical observations and the development of astronomical instruments

(iv) Adapt and expand existing visualization tools to the specific requirements of VirtU

2. VirtU will serve four distinct but overlapping astronomical communities: theorists, phenomenologists, observers and instrumentalists. Theorists will be able to exploit grid technology to carry out larger and better simulations. Phenomenologists will be able to access and analyze the largest simulations carried out by expert teams such as the Virgo consortium or members of the "UK Astrophysics Fluid Facility" (UKAFF) in a convenient and efficient manner. Observers and instrumentalists will be offered an unparallelled resource: extensive mock catalogues of realistic simulated objects which can be selected, "observed," and analyzed in a similar way to real objects. This resource will be invaluable for designing observing programmes and strategies, for interpreting observational data in terms of fundamental physical concepts, and for specifying requirements for novel astronomical instrumentation.

3. VirtU will exploit the investment made by PPARC in AstroGrid. Catalogues of simulated objects will be accessed using the extensive interface software that AstroGrid has developed for handling datasets of real astronomical objects. By bringing the best theoretical predictions directly into contact with observations, VirtU will maximize the utility of astronomical data. For example, VirtU will play a central role in the exploitation of surveys planned using facilities such as the UKIRT wide-field camera, VISTA, SCUBA, VIMOS, NGST, ALMA, NGST, etc.

4. The TVO will provide widespread access to state-of-the-art simulations. Performing such simulations is a highly technical and complex task, the preserve of a small number of specialist groups which possess the required expertise and access to expensive computing facilities. The TVO will make simulation data publicly available, thus allowing many groups to analyze them, greatly multiplying the scientific returns from the investment in simulation work.

5. The TOI will systematize and grid-enable diagnostic tools crucial for intermeshing theoretical models and observations, including libraries of synthetic galaxy spectra, statistical methods for parameter estimation, measures of spatial clustering, etc.

6. VirtU will deploy novel techniques for visualizing simulations and real data cubes, particularly stereoscopic techniques. Where they already exist, VirtU will extend and adapt techniques to the particular

requirements of simulations and modern data cubes and, where they do not, it will develop new techniques. The datasets making up the TVO will be huge so remote visualization tools will be essential.

7. The TVO will adapt and apply grid technology to enable simulations to be carried out at a variety of remote sites. Some of this work will be carried out through a collaboration with the Distributed European Infrastructure for Supercomputer Applications (DEISA), a pan-European collaboration of major supercomputer centers which includes the Virgo consortium. DEISA is currently seeking 14M euros of support from the EC.

8. The VirtU initiative is a collaboration involving most groups active in cosmological and gasdynamic simulation work in the UK. This includes the internationally known Virgo Consortium for supercomputer simulations, UKAFF, cosmologists and extragalactic observers, and computer scientists from EPCC and the University of Durham.

9. The techniques that VirtU will develop have applications across many other areas of e-science, including plasma physics, hydrodynamics, medical imaging and genomics. Some of its outputs, such as movies of cosmic evolution, have an enormous potential for public outreach and specialized educational programmes in schools and universities.

10. This is a 3-year programme for which we request funding for 9 code developers/computer scientists, a Project Manager/Senior Developer, 50% of a Project Scientist and hardware: a large disk storage facility to hold the TVO data, and **2** small-scale Beowulf clusters with large disk capacity.

2 The concept of VirtU

We propose to construct the Virtual Universe (VirtU). VirtU will be built upon a dynamic archive constructed, in the first instance, from state-of-the-art cosmological simulations which encode our current understanding of our world model, of the clustering evolution of dark matter and of the physics of galaxy formation. By including the latest, most realistic models of galaxies, clusters and other structures, together with relevant analysis tools, this virtual universe will be a powerful entity providing a service to a wide community of theorists, phenomenologists, observers and instrumentalists. VirtU, however, is not restricted to Cosmology. In the longer term, through the participation of UKAFF, it will encompass a wider range of simulations, covering, for example, star and planet formation. VirtU is an essential complement to AstroGrid and will form a key component of the Virtual Observatory (VO) by providing a twenty-first century computing infrastructure to enable direct and rigorous comparisons of the best theoretical modelling with real data. By enabling the link to be established between astronomical observations and astrophysical theory, VirtU will maximize the scientific returns from large observatories.

VirtU has two major components, the "Theoretical Virtual Observatory" (TVO), and a toolkit of applications targeted at the theory/observations interface (TOI). The relationship between them and with AstroGrid is illustrated in Fig. 1.

A. The Theoretical Virtual Observatory

The TVO is a completely novel concept. Member scientists, in close collaboration with Europe's Virtual Observatory programme, will build up the infrastructure required to publish simulated data and analysis tools in standardized formats. The best simulations will become readily accessible to non-specialists, leading to entirely new science applications. While the need for a Virtual Observatory has been universally recognized, the value of including theoretical simulations is not yet fully appreciated outside Europe. This collaboration and its European allies have a unique opportunity to become the primary driver in a cutting-edge activity which will develop very rapidly over the next few years.

B. The Theory/Observations Interface

To exploit the scientific opportunities offered by advances in simulation capability and in the quality of observational data, it is necessary to develop an interface to facilitate the comparison of models with data. This must be responsive to the limitations of both models and data and attempt to bring both on to a common plane by, for example, including noise or selection effects in the models or degrading the resolution of an observed image to the resolution of a simulation. TOI will develop this interface and make available on the grid a suite of diagnostics that include traditional as well as novel approaches to data compression, classification and parameter estimation in large datasets of galaxy spectra and images.

2.1 Scientific background and motivation

VirtU is the response of a large community of theorists, phenomenologists, observers and instrumentalists to the challenges presented by the ongoing explosion of observational and simulated data. This community wishes to have ready access to the latest simulations and models in order to further theoretical understanding, relate observations to basic theory and define the specification of new instrumentation by using models to anticipate the outcome of future observational programmes. This community also wishes to have equally easy access to diagnostic applications that are commonly used in the analysis of both data and simulations.

2.1.1 The basis for the TVO

Of all the branches of Physics, Cosmology is one which has experienced some of the most rapid progress in the past two decades. During this period, a comprehensive model of cosmic evolution has been formulated, developed and rigorously tested. This model, now widely accepted as the standard cosmogony, is based on two key assumptions: (i) that the Universe underwent an early period of inflationary expansion during

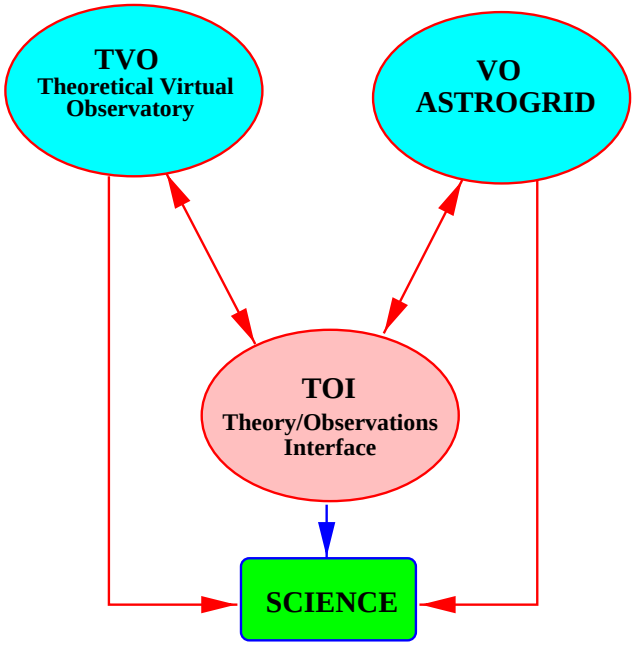

Figure 1: The relationship between the TVO, TOI and AstroGrid

which its curvature was flattened and small irregularities of quantum origin were imprinted and (ii) that these irregularities grew into cosmological structures by gravitational evolution driven by massive, weakly interacting elementary particles or cold dark matter (CDM). The CDM paradigm specifies the initial conditions for the formation of structure and makes definite predictions for, for example, the spectrum of temperature fluctuations in the microwave background radiation. This prediction has provided the most spectacular test of the theory so far, culminating with the recent, exquisite measurements by the WMAP satellite. These agree remarkable well with the paradigm and enable accurate estimates of some of the fundamental cosmological parameters. Although not all the parameters are constrained to the same high accuracy, the data indicate a flat universe in which baryons provide 4%, CDM 24%, and dark energy 72% of the cosmic energy density. This model, called ΛCDM, agrees not only with the structure of the microwave background radiation, but also with the distribution of galaxies, as mapped by the new generation of surveys, the Anglo-Australian "2-degree field" galaxy redshift survey (2dFGRS), the related quasar survey (2dFQRS) and the US-based Sloan Digital Sky Survey (SDSS).

The unambiguous prediction of initial conditions within the CDM paradigm makes it possible to calculate cosmic evolution. The early stages leading, for example, to temperature fluctuations in the background radiation at recombination, can be computed precisely using linear theory, but the late phases cannot be calculated analytically because the universe is strongly non-linear on small scales. The evolution of dark matter involves only gravitational interactions and, fortunately, there are accurate, efficient N-body methods to simulate it in a large computer. Recent N-body simulations by the Virgo consortium and others have established the distribution of cold dark matter on scales ranging from small galaxy halos to the Universe as a whole. This part of the problem may now be regarded as essentially solved (although the innermost density profiles of dark matter halos remain a subject of debate.) The part of the problem that remains unsolved is the formation and evolution of galaxies and related phenomena such as the nature of quasars and the evolution of intergalactic gas. Most current extragalactic observational programmes are targeted at these problems which provide much of the scientific rationale for large instrumental projects, from XMM/Newton to Gemini, VLT, VISTA, ALMA and NGST.

The astrophysics of galaxy and quasar formation are undoubtedly complex. They involve a wide range of processes and scales: gas cooling into dark matter halos, the formation of stars and central black holes, mergers of subgalactic fragments, galactic and extragalactic gas flows, energy injection by supernovae and AGN, and many more. Observational studies, on the other hand, are invariably subject to selection effects and suffer, in addition, from a fundamental limitation, the inability to associate objects seen at

a different epochs with a unique evolutionary sequence. Thus, for example, it is impossible to relate, from observations alone, the populations of SCUBA or Lyman-break galaxies seen at high redshift, with the population of galaxies in the universe today. Progress in understanding the formation and evolution of cosmic structure requires a close interplay between observations and detailed astrophysical modelling based on current cosmological theory. Facilitating such interplay is the overriding purpose of VirtU.

2.1.2 The need for TOI

Extragalactic Astronomy is undergoing a revolution. The exponential growth of observed and simulated data calls for radical new approaches for analysing and contrasting the observed universe with theoretical predictions. New surveys of large scale structure, such as the 2dFGRS, 2DFQRS and SDSS, have already produced redshifts, spectra and photometry for many hundreds of thousands of galaxies and tens of thousands of quasars. Ongoing and future projects include 2MASS/6dF, and a variety of surveys to be carried out using novel instruments such as VISTA, VIMOS, Deep2, VLT, Planck, etc. Many of these surveys may be viewed as laboratories for estimating cosmological parameters and for investigating galaxy formation and evolution.

There is currently a gap between the huge amount of data and the ability of the community to analyse them effectively. PPARC's e-science initiative provides a unique opportunity for a coordinated, systematic approach to the interpretation and analysis of the new datasets. VirtU will turn the currently cumbersome intercomparisons of different models with data into a straightforward e-science task, accessible to a large constituency. Many of the techniques to be developed here will be transferable to, and, more importantly, usable through AstroGrid.

2.2 The international context

Much current progress in Astronomy stems from collaborative work by partners across the globe. In order to enable these partnerships, suitable technologies are required to provide the transparent access to multisourced data and information sets. The VirtU initiative aims to ensure the prominent position of the UK at the international forefront in both the generation of, and the access to the best theoretical modelling and simulations in Astronomy.

Through the Virgo consortium, this application has strong international connections, particularly with cosmologists at the Max Planck Institute for Astrophysics in Germany, McMaster University in Canada and the University of Pittsburgh in the USA. Remote computing and access to large datasets are crucial for the successful and efficient operation of this collaboration. Virgo is one of the world leaders, if not the world leader, in the subject of cosmological simulations but there are several strongly competing groups in the world, located primarily in North America, Japan and Europe.

Partly as a result of the cohesiveness of the Virgo consortium, European cosmologists are the first to recognize the importance of, and make serious attempts to construct a theoretical counterpart to the Virtual Observatory. Work is already underway in Germany, as part of a high profile theoretical sub-programme of GAVO, the nationally funded German Virtual Observatory. This proposal has strong links with this effort through the Virgo consortium and through a Framework-6 Marie Curie Training Network proposal in which Garching and Durham are the major nodes. We intend to cooperate closely with our colleagues in Germany and also with Virgo colleagues in Canada who are involved in the "Sharcnet" grid project. VirtU will aim to ensure that interoperability issues with international partners are addressed, through e.g. the International Virtual Observatory Alliance (see section 6.5), thus enabling visibility and interaction with global theory-side resources.

VirtU members are fully engaged with developments in the wider 'grid' computing world, aimed at exploiting the new high speed networks coupled with distributed compute/storage clusters, to enable terascale supercomputing applications. The need to coordinate the development of computational capabilities across the European Community has been recognised both in the Community's Framework

Plans and in the long-term plans of the major European supercomputer centres. Computational Cosmology will play a significant role in these plans. Thus, the programme of the Virgo Consortium, the most visible transnational supercomputing consortium in European Astrophysics, has been adopted as a Joint Research Activity (JRA) by a group of seven leading supercomputer centres currently proposing for substantial Community support through the "Distributed European Infrastructure for Supercomputing Application" (DEISA) project (see Section 6.3). DEISA's goal is to deploy and operate a persistent distributed terascale facility to further computational science in Europe, and a key element of this application, the proposal to grid-enable simulation codes (Section 4.7.2), is intimately linked to our participation in the DEISA project. Indeed, in this application we are seeking resources to supplement those being sought from Brussels in order to cement our participation in DEISA.

Members of the US theory community are taking the first steps to integrate with the US National Virtual Observatory (NVO) project. The next NVO demonstration capability to be developed for the IAU General Assembly (July 2003) will focus on the mining of Globular Cluster Simulation Datasets aiming to identify globular clusters from observational sets, and comparing with a 4-D simulation of a globular cluster. It is of note, however, that theory activities within the NVO are of limited scale and that the US has, as yet, no major concerted effort on the scale of VirtU.

Computational Cosmology is a field where European Community scientists have had a major impact on the world scene. It is also a field which is strongly driven by computer technology and which is therefore developing extremely rapidly. Through its strong links with European colleagues, VirtU seeks to integrate the UK theory community more fully into current European VO programme and offers a unique opportunity for the UK to become the primary driver of a cutting-edge activity which will develop very rapidly over the next few years.

2.3 Outreach

Our proposed programme has ramifications beyond its main mission of providing a service to the astronomical community. Astrophysics and Cosmology stimulate the public imagination more than any other physical science because of the remoteness of the frontiers they explore, because of the beauty and immediacy of astronomical images, and because of the depth of the questions they address: how was our universe created? how will it end? what is it made of? where were the elements made? how were the Sun and the Earth born? how will they die? Cosmological simulations have traditionally played an important role in scientific outreach because they convey our current picture of cosmic creation and evolution both accurately and vividly. They are routinely used by museums, planetaria and popular science magazines to stimulate public interest in and awareness of science. For example, last year, the Virgo consortium was selected to participate in the Royal Society's most prestigious outreach activity, its Summer Science Exhibition. The Virgo exhibit which had as its centrepiece the stereoscopic movie "Universes to order", was voted one of the top three by the several thousand members of the public who attended.

The facilities provided by VirtU will be useful not only to professional astronomers, but also to students, particularly at University level, but also below. To take one example, appreciating the evolution of the universe of galaxies through visualization of simulations does not require much technical background and it is easy to imagine schools using such resources extensively. Similarly, the ability to compare simulations and observations on the Internet will undoubtedly give rise to numerous final-year undergraduate research projects in Physics departments across the country.

The spin-offs of our programme will extend beyond education and public outreach. Most of the statistical methods that we will implement in VirtU are generic and have applications in computer science, engineering, medical imaging, oceanography and other fields. As past experience demonstrates, methods developed for astronomical research find use in completely different areas and, in some cases, lead to commercial exploitation. The Maximum Entropy method is a good example of this.

3 The foundations of VirtU

VirtU is a national collaboration that includes the majority of the theory community in the UK involved in astrophysical simulations and their exploitation. The two main astrophysical simulation groups in the country, the Virgo consortium and UKAFF, are partners in this application. The third plank of the foundation for VirtU is provided by the emerging Virtual Observatory.

3.1 The Virgo consortium for cosmological simulations

Founded in 1996 as a "Grand Challenge" project, the Virgo consortium has a core membership of fifteen scientists in four countries, with major concentrations of activity and computing resources in Durham and Garching. C.S. Frenk is the PI in the UK and S. White in Germany. In the UK, the following Universities are represented in the core group: Edinburgh (Peacock), Cambridge (Efstathiou), Durham (Frenk, Jenkins, Theuns), Sussex (Thomas), Nottingham (Pearce). In addition, there are core members in Germany, Canada and the USA. At any given time, an additional 20-25 scientists, mostly PhD students and postdocs, are directly involved in aspects of the Virgo programme. Virgo has access to large supercomputing resources in the UK and Germany, most notably the dedicated 152-processor "Cosmology Machine" at the Institute for Computational Cosmology (ICC) at Durham and the IBM-Power4 768-processor Regatta of the Max-Planck Rechenzentrum in Garching.

Since 1998 core Virgo members and colleagues have published over 50 papers based on Virgo simulations. These include some of the largest and best simulations over the entire range of cosmological problems, from the distribution of dark matter on large scales to the innermost structure of galactic halos, and from the properties of the intergalactic medium at high redshift to detailed studies of the intracluster medium. Recent highlights include the 1-billion particle "Hubble Volume" simulations of dark matter over virtually the entire visible universe and their use to construct realistic mock catalogues of the 2dF galaxy and quasar redshift surveys which have been instrumental in the exploitation of these unique datasets; the discovery of a universal mass function of dark matter halos in hierarchical clustering theories; the first N-body/gasdynamical simulation to produce a population of simulated galaxies with a luminosity function and clustering properties in approximate agreement with the data.

The Virgo Consortium is the best-known cosmological supercomputing consortium in the world. This is due not only to the quality of its simulations, but also to its aggressive policy of making images, movies, simulation data and numerical codes readily available to the international community. Much of the Virgo simulation data can be downloaded on the Internet (http://www.mpa-garching.mpg.de/Virgo/). Virgo images (available at the Virgo download site in Garching and also at http://www.virgo.dur.ac.uk/index.html turn up in magazines all over the world, and Virgo numerical data have been used for many research papers that the consortium members found out about only after they were published. Virgo codes are being included as example science applications in the grid-computing proposal recently submitted by European supercomputer centres for EC support under the DEISA programme.

3.2 UKAFF

The UK Astrophysical Fluids Facility (UKAFF) was established through a PPARC submission to the Joint Research Enterprise Initiative (1999) on behalf of a consortium of 23 university astrophysical theory groups. UKAFF's remit is to provide supercomputing facilities for astrophysical fluid calculations, with peer-reviewed access open to the entire UK theory community, and to ensure that expertise in computational astrophysical fluid dynamics is spread as widely as possible within the UK. The distinctive feature unique to UKAFF is that there is no privileged access to it: all time is allocated solely on the basis of peer review, and any UK institution is eligible, just as for a telescope.

The JREI grant of 2.41M allowed the purchase of a 128 CPU SGI Origin 3000, the first to be installed in Europe. The facility also operates a smaller 24 CPU Origin 2000 for development work. UKAFF runs

regular parallel programming training courses (which are invariably oversubscribed) to ensure that the community is trained with the skills required to use the facility.

Since UKAFF began operating in full production mode on 1st January 2001 over 60 publications have resulted from simulations performed at the facility. UKAFF's success has resulted in PPARC inviting an application from the facility for a major upgrade to the computer systems. The University of Leicester, where the facility is housed, is also committing funds from SRIF-2 to provide significant additional machine room space to house the upgrade.

3.3 Diagnostic Tools

The applicants in this proposal include theorists who have featured prominently, and even led some of the major extragalactic surveys in which the UK has been involved over the past two decades. Examples are the Durham/AAT galaxy survey, the QDOT and PSCz surveys of IRAS galaxies, the APM survey and the 2dFGRS. As a result, this group of applicants has vast expertise in designing and implementing techniques for extracting information from datasets that makes sense from a theoretical point of view and, of course, in relating this information to theory. Examples include techniques for characterizing the clustering pattern of galaxies which have been adopted by groups worldwide and methods for extracting physical parameters from galaxy spectra. The collective expertise of this group is unrivalled in the world and will be crucial for the successful implementation of TOI.

3.4 The Emerging Virtual Observatory

The Virtual Observatory (VO) is a federative project which has independently been accorded high strategic priority by the astrophysical communities in a significant number of astronomically active nations across the globe (e.g. the USA, Canada, the UK, the Netherlands, Germany, India, ...). These efforts have rapidly coalesced into a coordinated approach to building a Global Virtual Observatory. This will be a distributed web- or grid-like computational structure which links the data archives of all major astronomical telescopes, both ground- and space-based. A user anywhere in the world will be able to formulate queries which require the location and interrogation of disparate and distributed archives without any detailed knowledge of the existence or properties of those archives. While almost all VO work is currently directed towards observational datasets, there is also considerable demand for simulation datasets which are essential for interpretation of the observations. This is demonstrated by the many Virgo images and the many independent analyses of Virgo simulation data which have appeared in popular and professional journals. Virgo-developed simulation codes are also in regular use at many institutions worldwide that have no direct relation to the consortium. However, Virgo's VO functionality at present is rudimentary, with extremely limited access to data and information. (eg see the current Virgo down-load site at http://www.mpa-garching.mpg.de/Virgo/).

AstroGrid is a £3.7M project (http://www.astrogrid.org/) which is developing a working datagrid for key selected databases, with associated data mining facilities, with rollout iterations through to late 2004. It is part of the worldwide drive towards the VO concept, and can be seen as the UK's contribution to this vision. However, it is both wider and more focused than other initiatives. It is wider in that it covers Astronomy, solar Physics, and space plasma (solar terrestrial) Physics, and covers all wavelengths from radio to X-ray. The project is also part of a coherent UK e-science programme, with links to projects in particle Physics, bio-informatics, and basic grid technology development.

AstroGrid is, however, focused in that it is developing a fully functional (but limited in scope) VO on a short timescale, to enable early scientific exploitation of the system, and provide technological lessons to better inform the deployment of larger scale VO systems. This is requiring a concentration on selected datasets. Thus AstroGrid's priority is to develop a virtual observatory capability to support efficient and effective exploitation of key astronomical datasets of particular importance to the UK community, for example data from WFCAM, VISTA, XMM-SSC, e-MERLIN, SOHO and Cluster. It seems clear that

good data curation, archive management, and data mining services all need to be closely linked together. AstroGrid is therefore a partnership formed by UK archive centres and astronomical computer scientists.

AstroGrid is an integral part of the Astrophysical Virtual Observatory, an EU funded initiative with partners including ESO and the CDS in Strasbourg. It has also taken a leading role in the formation of the International Virtual Observatory Alliance (IVOA, see Section 6.5), where it leads the development of key interoperability standards.

AstroGrid offers a strong infrastructural basis upon which VirtU can prosper and develop key theory-side services. It provides a crucial link to observational datasets and, at the same time, a pervasive delivery mechanism for the synthetic data exposed by VirtU. The interface programme is described in section 4.4

4 The structure of VirtU

We now turn to a detailed description of the elements that together will provide the essential infrastructure for VirtU. Figure 2 illustrates their interconnection in a schematic way.

4.1 Construction of the TVO

In the first instance, the TVO will be built around the **Millennium** simulation, the largest and most ambitious cosmological simulation ever conceived, currently being planned by the Virgo Consortium. Virgo is already responsible for a previous record simulation, the *Hubble Volume* calculations which were carried out in 1998. These followed the evolution of cosmic structure in a cube of side 4 Gpc using 1 billion dark matter particles. Because of their sheer size, these N-body simulations set a milestone in the field of computational physics. More recently, Ostriker and collaborators have carried out even larger N-body simulations, with 2 billion particles. The Hubble Volume and related, smaller simulations by Virgo were made available to the Astrophysics community via the Virgo website (see Frenk et al "Public Release of N-body simulation and related data by the Virgo consortium," astro-ph/0007362) and these data have, indeed, been exploited by many groups around the world. In total, over two dozen papers have been written by groups which are *not* part of the Virgo collaboration using Virgo data. The public release in the pre-grid era of datasets which were enormous by the standards of the day (tens of Gbytes) proved to be a challenging task from which we learned valuable lessons.

• The millennium simulation: a foundation for the TVO

The Millennium simulation will follow 10^{10} particles in a periodic volume of a ΛCDM universe 500 Mpc on a side, with cosmological parameters as determined by WMAP and 2dFGRS. This simulation will have 10 times more particles and 1000 times better mass resolution than the Hubble Volume runs. It will track the clustering evolution of dark matter and the formation history of about 50 million galaxies. Data will be output about 50 times (enough to construct lightcones and make movies) and each output will consist of 300 GByte. Thus, the raw data alone will amount to 15 Tbytes, *twice the size of the current HST data archive!* The simulation will be carried out using an optimized version of the new, highly efficient *GADGET-2* code (based on the tree-PM method) which can execute N-body simulations of unprecedented size on the newest generation of supercomputers. The simulation will be performed on the IBM-Power4 768-processor Regatta of the Max-Planck Rechenzentrum in Garching. This is one of a handful of supercomputers in Europe which can currently provide the required 1 Tbyte of memory. This calculation will set a new milestone in computational physics and redefine the state-of-the-art in cosmological simulations for years to come.

The Millennium simulation will be populated with galaxies using the techniques described below. The outcome will be a virtual universe containing a full description of the visible properties of each galaxy *at all epochs*, such as spectral properties from the far-UV to the submillimetre, gross morphological characteristics, star formation rates, gas content, and, of course, spatial distribution. The Millennium simulation populated with galaxies constitutes the minimum specification for the TVO. However, as discussed in Section 4.6, the TVO is intended as a dynamic entity which will grow continuously.

The Flow of VirtU

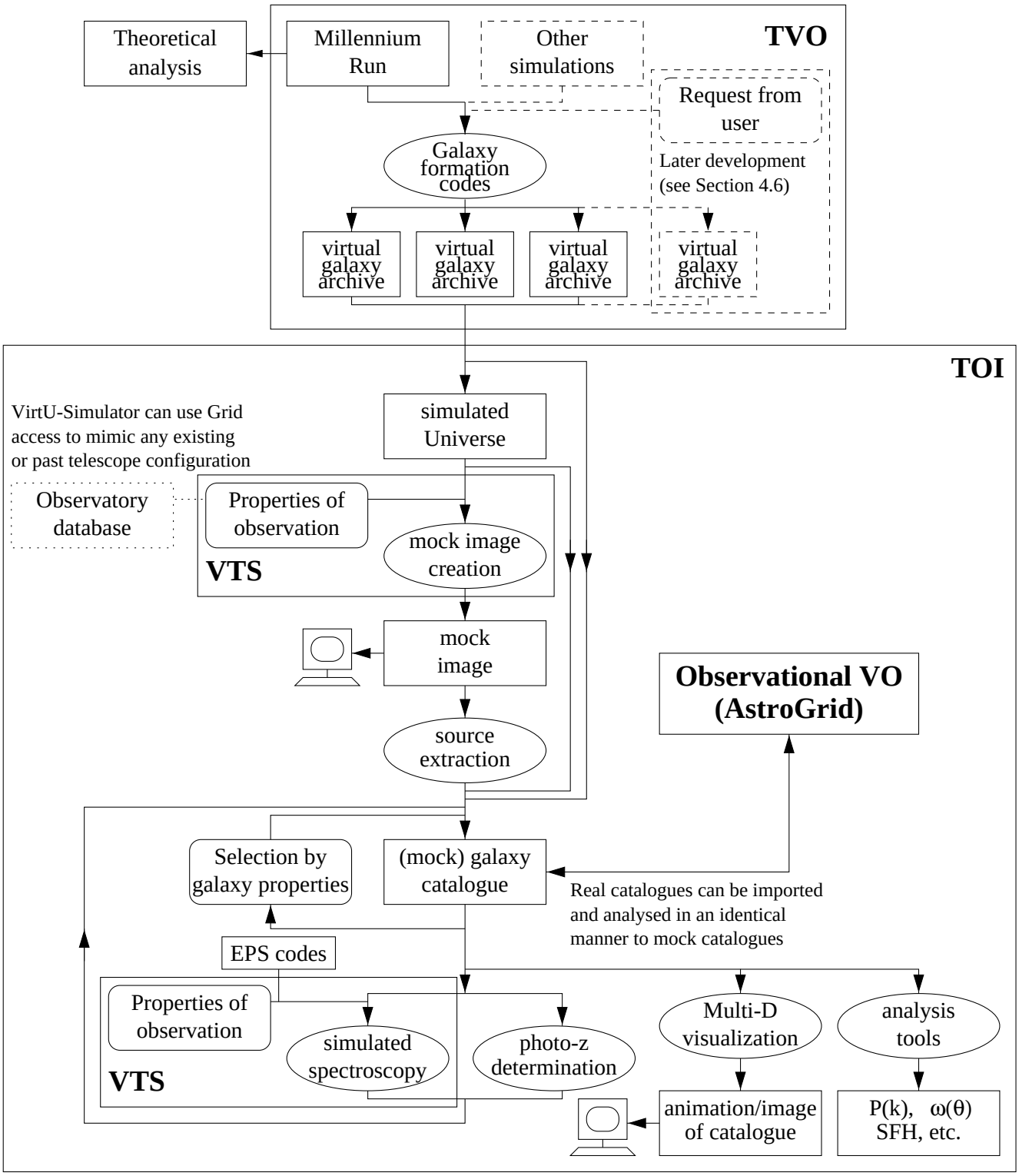

Figure 2: An example of an extragalactic application of VirtU. A similar structure can be used for other applications, e.g. star-formation

4.2 Metadata and data models for synthetic data

A key problem identified at an early stage by the observationally focused VO project, is the issue of standardised data and metadata models to describe the data resources that can populate the VO. Thus, data objects must be described in a manner such that computer agents can manipulate the relevant objects. The focus here is on semantics.

Astronomy has developed a flexible file transport format, the widely used FITS. This format has a limited concept of metadata, with FITS KEYWORDS being used to describe the data. However, FITS KEYWORDS have the significant limitation that they are not unique. For instance, two observatories could produce 'V'-band data, but in fact the filter transmissions of the two could be very different. Thus, computer assisted comparisons of the two sets could be spurious.

Significant effort is now underway to provide a semantic infrastructure (e.g. Williams 2003) to enable the effective description and access to observational datasets. The community - through the IVOA (see section 6.5) - is working to develop standards in a number of areas, including:

- Data Access Layer: defining and formulating VO standards for remote data access. Client data analysis software will use these services to access data via the VO framework; data providers will implement these services to publish data to the VO.
- Data Model: the abstract representations of astronomical data, e.g. space-time coordinate metadata models.
- Registry: addressing the need for an astronomer or computer process to be able to locate, get details of, and make use of any resource located anywhere in the Virtual Observatory. Individual resource registries must be locatable and query-able in some standard fashion.
- Content description: unique descriptors for astronomical data, e.g. POS_EQ_RA_MAIN - a right ascension.
- VOTable: an XML standard - now in place - to represent tabular data, but extended to represent a wider range of data.

VirtU will participate in these standards processes to ensure that the requirements demanded by simulation datasets are addressed. It is worth noting that the IVOA working groups are at the early stages of developing standards, and that VirtU will be able to inject significant input from the project start in mid 2004. Many issues requiring solution will be similar to those for observational datasets. However, there are a number of specific demands set by the theory community. One of these is the handling of 'point' or 'particle' datasets. Much model data are generated in the form of particles, each having an N-dimensional set of attributes (mass, chemical composition, position, etc). A generalised data structure with associated metadata will be required to accommodate these attributes and allow efficient data exchange. Thus, a key early deliverable for VirtU will be the deployment and uptake of a unified data format or model data.

4.3 Synthetic data access

VirtU, as well as offering access to archived raw simulation data, will provide refined data products such as group catalogues, merger trees and galaxy catalogues derived from the raw data. The ability to create and manipulate these data products is an essential functionality needed both to build and update the TVO. In the first instance, the TVO will contain the raw data from the Millennium simulation and other suitable dark matter N-body simulations. Developing a software pipeline which will allow the incorporation of datasets such as these, and eventually other simulations, into the TVO is a high priority for our programme.

• Mock galaxy catalogues

We propose to employ at least three different techniques to populate the Millennium and other N-body simulations with galaxies. The first of these, which is rapidly becoming the standard method for following the evolution of the galaxy population, is semi-analytic modelling of galaxy formation. This is a physically

based approach whose development has been led by a subset of the applicants at Durham and Garching. The other two techniques are empirical. In one of them (the *high-peak bias model*), dark matter particles are tagged as "galaxies" according to a heuristic "biasing" scheme which depends upon the initial, locally smoothed dark matter density field. In the other (the *semi-observational model*), simulations like the Millennium are populated with galaxies by making a correspondence between dark matter halos ranked by mass and real galaxies ranked by luminosity.

From the e-science point of view, populating the Millennium simulation with galaxies using the semi-analytic method poses a particular challenge. Galaxies form from gas which cools and condenses inside dark matter halos which, in turn, grow by accretion and mergers. The Millennium simulation will be able to resolve the merger trees and follow the growth of about 50 million galaxies. The merger trees need to be reconstructed from the 15 Tbyte dataset by identifying halos and tracking their evolution from one output to another. Virgo members have already developed a state-of-the-art parallel substructure finder, but, as discussed in Section 4.7, adapting it to a dataset this size is non-trivial. Once a merger tree has been constructed, galaxy formation is followed, starting from the upper branches of the halo merger tree which correspond to the lowest mass progenitors of the final halo. The semi-analytic model calculates the formation of the visible galaxy using sophisticated models to describe the cooling of gas, its fragmentation into stars, the generation of metals and dust, feedback processes, galaxy mergers and the evolution of stellar populations. In this way, the physics involved in the build-up of galaxies are followed in detail. Again, although the code for this (in fact two independent codes, one from Durham and the other from Garching) exists it has never before been used on a dataset this size.

The heuristic biasing schemes that we will apply were introduced in the 1980s. Pioneering simulations of structure formation in a CDM universe carried out by some of the applicants in this proposal suggested that galaxy clustering could be explained if the galaxy distribution was actually a spatially modulated form of the dark matter clustering pattern. In this "high-peak" biasing scheme, the probability of finding a galaxy is related to the initial smooth density field. This, and derivative schemes, have been implemented by the Durham group in the Hubble Volume simulations to make mock catalogues of the 2dFGRS that have been instrumental in the tremendous impact that this survey has had in studies of the large scale structure of the Universe. Finally, the high resolution attainable in modern N-body simulations makes it feasible to implement a third, "semi-observational," scheme to populate a simulation with galaxies. In this scheme, halos and their subhalos are ranked by mass and assigned a "galaxy" whose properties are drawn from observed galaxies of the same spatial abundance in a survey such as SDSS.

The generation of a library of galaxy catalogues drawn from the Millennium simulation is described in the Project Plan. The catalogues from the Millennium simulation providing a complete description of the evolution of the galaxy population will be available for "observation" by virtual telescopes within the TVO. This capability is critical for understanding how different observing strategies constrain the uncertain physical assumptions which underlie the simulated catalogues. The network's TVO will become an essential tool for the physical interpretation of major observational programmes such as the ESO's VIMOS, Keck's DEIMOS-DEEP, or the 2dFGRS and SDSS surveys, greatly enhancing their scientific value. For example, by providing an evolutionary timeline for each model galaxy, the TVO will enable studies of the relation between different classes of galaxies observed at different cosmic epochs.

4.4 The VirtU-VO interface

VirtU, as we have explained, has two major strands: the construction of a Theoretical Virtual Observatory populated with model data, and the development of an infrastructure for confronting theoretical models with observational datasets. Section 2.1 detailed a number of key science drivers involving direct comparisons of data from observations with predictions from simulations. The interface of VirtU with the developing observational VO systems will be critically important in allowing the interaction of theory and observational resources.

In practical terms, VirtU will work closely with AstroGrid to ensure the interoperability of services. The AstroGrid architecture is based on the concept of configurable components which together form the VO.

This service-based architecture approach is outlined in the relevant section of the AstroGrid Phase A report at http://wiki.astrogrid.org/bin/view/Astrogrid/RbArchitectureOverview. VirtU will develop its theory-side components using AstroGrid Application Programme Interfaces where appropriate.

In Section 6.1, an expanded description is given of how VirtU will deploy its infrastructure based on that currently under development by the AstroGrid. In coordination with AstroGrid and partners in Germany, especially GAVO, VirtU will aim to ensure the development and deployment of appropriate interfaces into the wider Euro-VO infrastructure. This process is somewhat simplified by the fact that AstroGrid has taken a lead role in technology development of the Euro-VO. Thus, the interface to AstroGrid will be similar, if not the same, as the interface to Euro-VO.

4.5 The VirtU-Data Centre interface

4.5.1 Data storage and retrieval

There are several groups around the world which specialize in performing large simulations. One of the goals of VirtU is to design and implement the infrastructure required to make it possible for a large and disparate collection of simulations to be incorporated into the TVO. A related goal is to make these data accessible to individuals around the world who, in most instances, will only have access to computers that are orders of magnitudes smaller than those in which the simulation was performed.

The development of tools for accessing simulation data produced on computers with different architectures using a variety of codes, poses an interesting e-science challenge. For example, at present, it takes a considerable amount of effort simply to read the often curious data formats that different simulators use. Even more daunting is the task of unravelling the physical units of the variables in different simulations. An observer with new, exciting data may be reluctant to compare against theoretical models derived from 'black box' simulations where it is not clear whether the theoretical prediction falls within the range of validity of the simulation and the underlying models.

A good example is provided by the *Santa Barbara Cluster Comparison* project (Frenk et al. 1999), an international collaboration organized by two of the present applicants in which almost all practitioners of cosmological simulations worldwide took part. Even though precise instructions had been issued to the simulators, a disproportionate amount of effort was required simply to compare the various simulation outputs and many mistakes were found, often due to misinterpretation of code units. This example shows how important it is to develop a foolproof infrastructure if people who are not familiar with the codes are to make use of the simulations. It also illustrates the scale of the problem: far from just serving the data, this infrastructure should permit *quality control* of the simulation itself and also of the results of a query. The user of an archive may be happy just to obtain a dark matter catalogue from a simulation, but should, when required, be able also to check the gravitational softening, timestep, integration scheme and all such coding details which may be relevant for certain types of query. Just like the VO needs to archive the raw data and calibration frames in order to allow users to check the result of a query involving derivative data products, so VirtU will need to allow the archiving of raw data and tests to allow the user to verify query results. The organisation of such a system is essential for a correct usage of simulations and is ideally suited for an e-science initiative.

The problem of intercomparing simulation results is magnified when different simulation techniques are involved. Consider, for example, comparing the output of a "Smooth Particle Hydrodynamics" (SPH) code, which consists of particle positions, velocities and thermodynamic properties, with output from a Eulerian mesh code, which consists of data on a grid, or output from an Adaptive Mesh Refinement (AMR) code, in which the data are given at the vertices of a system of interlocking grids (see Section 4.6). Even the simplest comparisons, for example, of the gas density profiles of a bound object, require a considerable, time-consuming programming effort to bring these disparate outputs onto a common footing.

VirtU does not aim to enable every user to perform all possible tasks through a query, but rather it aims to provide tools for efficient data access. A user who has developed a new tool for, say, identifying dark halos in a simulation, should be able to access the raw data and apply this new group finder, without having

to worry about how each different simulation stores particle positions. A more demanding problem is the computation of, for example, mock quasar absorption spectra. A user may have written a proprietary code for computing mock spectra, and may want to try this on dark matter only simulations, as well as on SPH or AMR simulations that include gas. The required input data (HI densities, temperatures and velocities) are stored in very different ways in each of the corresponding data files. For example, for the dark matter only simulation, a prescription for estimating density and temperature would be required first. For SPH and AMR simulations, the interpolation schemes used to infer thermodynamic quantities are very different. Efficient tools for data access to the TVO should address these problems in a manner that is transparent to the user. Only then will it become possible to query a large number of different simulations to test for convergence, for example, and build up confidence in numerical predictions.

For the TVO to be viable, it is essential that we develop a framework to simplify and standardize the generation, storage, access and analysis of simulation data. Specific tasks of VirtU include:

(1) Design of tools to archive simulation data

These should address the issues of data quality, data locality, and access rights. Data quality requires the validation of the simulation code and of the chosen runtime simulation parameters, both those that affect the accuracy of the numerical integration (e.g. timestep, softening, tree opening criterion, etc) and those that affect the physics being modelled (e.g. resolution, gas processes, etc). Data locality refers to the actual location of the TVO data. Small simulations could be stored locally, but larger ones, like the Millennium simulation, will necessarily have to remain in one of the large data storage facilities on the computational grid of the TVO. Data access refers to different levels of authorization. For example, older or smaller simulations in the TVO may be universally accessible, but ongoing or very new simulations may only be accessible to members of a specific collaboration. The TVO tools should allow the correct implementation of access rights.

(2) Design of tools to query simulation data

The TVO will require the adaptation of existing tools and the development of tailor-made ones to query the simulation archive in a manner that is independent of the actual data structure in which the simulation data is held and the computer hardware in which the simulation was performed or the query launched. Some of the requisite tools already exist. For example, the hdf5 file format already guarantees (like FITS) that files can be shared across different computer architectures. But this only allows the user to read the files; more sophisticated tools will need to be developed such as a metalanguage to interpret the information contained in the files. If the query is computationally demanding, it should be possible for the query to be performed remotely on a larger authorized computer. The tools should also be able to select the optimal location for archiving the result of a large query and for recognizing repeat queries.

4.5.2 The virtual galaxy archive

The design of an efficient archive of galaxy properties is essential for a successful TVO. The galaxy archive will need to be browsed, filtered and manipulated in a number of ways in order to mimic different types of observation. A flexible, cross-cutting and readily accessible data store is the required starting point.

The primary data resource will be a galaxy data cube containing a variety of information. This is a store of basic properties for the galaxies present at each output time in, in the first instance, the Millennium simulation, but eventually also in other suitable N-body simulations. For semi-analytic models from the Millennium simulation, these basic properties include the mass of the dark matter halo in which the galaxy resides, the stellar mass and scalelength of the disk and bulge components of the galaxy, the mass of gas, dust and metals in the hot and cold phases and so on. We estimate that the Millennium simulation will contain 1 billion galaxies and about 30 numbers will be stored for each of them to specify its fundamental properties. In addition to this, we aim to be able to provide the luminosity of a galaxy at any specified wavelength or in any requested passband. The most general and future-proof way to achieve this is to store the full star formation history of each galaxy, allowing the construction of its full spectral energy distribution. The star formation history typically consists of the star formation rate

averaged over all of the galaxy's progenitors, divided into 10 bins of metallicity over 150 timesteps, plus details of any bursts of star formation that occurred, giving on the order of 2000 numbers in total for each galaxy, or 8 Kbytes. For one Millennium output slice with around 50 million dark matter halos, each containing, on average, one galaxy, this requires 400 Gbytes of data. For 50 epochs, each galaxy model will generate an archive approaching 10 Tbytes in size.

The data archive will be designed to permit the fast and efficient execution of a number of disparate operations which will enable the comparison of the TVO model predictions with different types of observation. For example, a user may want to examine the clustering of galaxies at a particular redshift, accessing a relatively small locus in the data cube or, instead, may wish to apply a photometric selection to the entire data cube and extract galaxies over some interval of redshift. The design and implementation of the TVO galaxy data archive underpins much of the subsequent activity of the TVO and is a critical element of the component of the Project Plan dealing with data formats and management.

In the first phase of operation, the user of the TVO will be able to select from a small library of pre-existing semi-analytic models, constructed by varying a few parameters, for example, those that govern the star formation or galaxy merger rates, and also from libraries built using the empirical galaxy-building methods just described, with different parameter values. Towards the end of the project, in a second, more ambitious phase of operation, the user of the TVO will be able to demand a different virtual universe altogether by invoking a particular variant of the semi-analytic model or of the empirical galaxy schemes. This functionality will depend upon the development of techniques to grid-enable the semi-analytic and empirical galaxy-building codes, as discussed in Section 4.7. In this mode, the user may generate a new model with a particular choice of parameters or may alter an entire module of the code, for example, the star formation law. The construction of galaxy catalogues from the Millennium simulation present a new set of challenges with which our existing software tools are ill-equipped to deal. Distributed operation of grid-enabled codes will be essential.

4.6 Extending and updating the simulation base of the TVO

So far, we have focussed on the inclusion, processing and manipulation of large N-body simulations in the TVO, in particular the Millennium simulation which represents the minimum specification for the TVO. However, the full power of the TVO will only be realized when a variety of simulations of different kinds are included. Specific aspects of the galaxy population can be simulated directly using N-body/gasdynamic techniques. These are much more expensive than semi-analytic calculations and cannot therefore be used to simulate large galaxy populations in detail but simulations targeted at specific problems such as galactic structure or the X-ray properties of the intracluster gas are extremely important.

Gasdynamical simulations in which the plasma is assumed to be a non-radiative ideal gas are relatively inexpensive and cosmological SPH simulations have been performed with up to 512^3 dark matter and 512^3 gas particles, while simulations of single objects like galaxy clusters, have been performed with about 1 million particles. Such simulations have been instrumental, for example, in studies of the X-ray emitting intracluster plasma and are the ideal means to relate theory to observations. In the first instance, the TVO will contain the large library of non-radiative gas cosmological simulations already built up by Virgo and by other applicants, in particular Ostriker and collaborators, and Bryan and collaborators. This library contains the best simulations ever performed in this area. It will be of tremendous value for the interpretation of data from X-ray missions such as Chandra and XMM/Newton. For example, in a PPARC-funded programme, researchers at Sussex are calculating the X-ray properties of the gas in their cluster simulations and casting them in terms of the quantities that can actually be measured by XMM/Newton, such as images, temperature maps and X-ray spectra.

The methodology being developed at Sussex can, in principle, be applied to any of the gasdynamical cluster simulations in the TVO. The TVO user will be able to dial up a cluster of given X-ray luminosity and "observe" it in the same way as a real cluster is observed with XMM/Newton, possibly including instrumental details such as integration time (see Section 5.1 for further discussion). Of course, the TVO user will know much more about the simulated cluster than the real one, for example, its mass, density profile, and formation history.

One area of cosmological research which has been revolutionized by gasdynamic simulations is the study of the gas clouds at high redshift that produce the Lyman-α forest seen in QSO spectra. The first credible simulations of the distribution of gas in a CDM universe, published in the mid 1990s, completely overturned the established view of the nature of these clouds and their role in galaxy formation. Such simulations are more complicated than simulations of the X-ray gas in clusters because they must treat the radiative cooling of the gas but, since the relevant gas overdensities are small, the problem is still tractable. In the past few years, the size and sophistication of these simulations has increased greatly to the extent that observers now routinely use simulations to interpret their data. This area illustrates better than any other how the symbiosis between theory and observations is the key to scientific progress in many branches of Astrophysics. The TVO will include the best current simulations of the high redshift gas together with the substantial body of software that has been written to extract and analyze observables, such as synthetic spectra, which can be directly compared with observations.

Gasdynamical simulations of galaxy formation are much more expensive and complex than simulations in which the gas is approximated as non-radiative. In this case, it is essential to include radiative heating and cooling, star formation, energy and heavy element injection from stellar winds and supernovae, and a physically consistent treatment of turbulent, multi-phase gas. The most ambitious and sophisticated simulations of this kind to date have been carried out by Virgo members and collaborators in the US and have pushed the gas and dark matter particle numbers in SPH simulations to 2×324^3. These simulations will form the nucleus of a growing sector of the TVO which will eventually enable detailed comparisons of, for example, galactic structure, with observations.

The TVO will be constructed with expansion in mind. Its structure will be sufficiently flexible that newer and better simulations will be readily incorporated as they become available ensuring that the TVO is a dynamic, evolving entity that contains the best theoretical models at any time. At a later stage, it may even become possible for the TVO user to perform a new simulation remotely on demand, specifying initial conditions and simulation parameters. The development of the appropriate tools for such higher order functionality is not included in this proposal.

4.7 The VirtU compute grid

4.7.1 Processing raw simulation data

The manipulation and processing of the Millennium data, in particular, poses significant e-science challenges. To process the Millennium simulation, given current network speeds, it will be necessary for the data and compute power to be local, but for other datasets it would be wise to prepare for the added complexity which arises when the data and the computing resources reside in machines sited at different geographical locations. The DEISA initiative described below will address aspects of this problem.

Members of the Virgo consortium have already developed an efficient parallel code to find groups in clustered data and to identify their substructure. (Only two such codes exist worldwide.) However, this code has never been applied to a dataset the size of the Millennium simulation. To do so poses a considerable e-science challenge because the memory requirement exceeds even that needed to perform the simulation in the first place. One of the main tasks in the Project Plan is to design a strategy to accomplish this. It may be that this will require dividing a raw simulation output into a number of spatial domains (with overlaps), each of a size that can be handled by a single parallel process. To process an entire output, a task farm procedure will need to be developed to divide the simulation output into manageable pieces which can then be distributed amongst individual processors and to combine the subcatalogues returned by the individual processors.

The next stage in making merger trees requires that group catalogues be available at every output (and that their spacing be sufficiently close). The merger trees are built by first identifying all halos at the present epoch. For each such halo, its progenitors at earlier times are identified by matching particle lists so as to create a tree structure that describes the entire assembly history of the halo by mergers of sub-units and stores the positions, velocities and other properties of each sub-halo. This requires

access to both the raw data and the group catalogues at all redshifts. In principle, this operation can be parallelised by means of a task farm in which each individual halo is handled by a single processor using a serial algorithm. However, the complexity is greater: unlike in the construction of the group catalogues in which each processor can proceed in isolation without the need for I/O, building an individual merger history requires each process to access the entire group catalogue at all epochs.

4.7.2 Grid-enabled distributed computing

The increasing availability of grid-enabled computer systems offers the potential for running a variety of modelling codes on computers at distributed locations. There are a number of tools, such as Globus, which control access to the grid. However, the modelling codes themselves need to have some level of "grid awareness". In this subsection, we describe our plans to grid-enable *GADGET* to perform simulations of galaxy formation, as well as some of the most widely used codes in UKAFF.

The two most obvious difficulties involved in grid-enabling a code are ensuring the portability of both data and code. The first problem will be overcome by developing a standard file format, as described in section 4.5.1. This will require some effort to adapt existing codes to use this standard format. The second problem, code portability, is relatively easy to solve but some effort is required to ensure that codes are efficient on a variety of operating systems and compilers. Thus, the requirements for grid-enabling our simulation codes are:

- Implement data file formats compatible across nodes.
- Develop scripts to compile the code at different nodes. Check that the code works consistently across nodes.
- Develop a monitoring system to check resource availability at different nodes.
- Develop a job management system to submit simulations at different nodes and to collate output at a single node.

4.7.3 Gadget applied to simulations of galaxy formation

As described earlier in this application, the last few years have seen a massive increase in the fidelity of dark matter simulations. The evolution of the dark halos that dominate the masses of luminous galaxies and clusters can now be simulated almost exactly using supercomputers. However, our understanding of the gas dynamical and feedback processes that structure the visible parts of galaxies is still rudimentary. As an example, consider the formation of galactic discs. About 70% of all galaxies have discs of gas and stars, and a sizeable fraction of late type galaxies are almost entirely disc-dominated. Most of the cosmic star formation at the present day is occurring in discs. Evidently, understanding the origin of discs is essential if we are to understand the present day morphologies of galaxies and may be a critical factor in determining the star formation history of the Universe. Yet, the formation of disc galaxies is poorly understood within the context of CDM cosmology.

According to simple models, disc galaxies form as a consequence of the collapse of cooling baryonic material within dark halos. The characteristic sizes and angular momenta of discs can be explained in this picture if the gas conserves it angular momentum during collapse. However, recent SPH simulations have demonstrated that the merging of gaseous subclumps results in a catastrophic loss of angular momentum of the baryonic component. Within the past two years, some of the groups represented in this proposal have began to achieve real progress in solving this fundamental problem. The key lies in modelling the process of feedback on the star formation rate by the injection of energy liberated during the course of stellar evolution. This requires simulating many galaxies at very high resolution.

To see how the Grid can make a real impact on this work, it is important to understand the difference between the simulations discussed here and large-scale dark matter simulations (*e.g.* the Millennium simulation) or classical grid-based hydrodynamic simulations. The latter computations are usually memory limited, so a state-of-the-art grid-based hydrodynamic simulation requires a supercomputer employing

the largest possible grid compatible with cpu cycle limitations. In contrast, the SPH simulations of individual galaxies have (by modern standards) negligible memory requirements but require many cpu cycles. Furthermore, since we wish to test and understand complex physics (which inevitably means parameterisations), we need to run many simulations. It would be possible to achieve rapid progress in this field, with the prospect of high scientific impact, by developing a grid based interface for distributed computing that will permit many simulations to galaxy formation with different parameters to be carried out across many small clusters at different sites. Since each simulation has a small number of particles (less than a million at this stage) only small quantities of data need to be transferred between nodes. The best code available for galaxy simulations is *GADGET*, written by Virgo members Volker Springel and Simon White, and this is one of the codes that we propose to grid-enable.

4.7.4 UKAFF codes

In some sectors of the scientific community, codes are provided centrally by a single organisation which will, in time, produce a grid-capable version of the code. Experience with UKAFF demonstrates that this is not the case for those modelling astrophysical fluids. Whilst several of the codes have a common root, they have evolved along different paths over several years. Each individual code needs to be grid-enabled. Within the scope of this proposal, we cannot achieve this for all codes. Instead, we aim to put in place the requisite tools and demonstrate their effectiveness by applying them to a small number of the more widely used codes. We will select a number of codes representative of those that are currently run on UKAFF for this pump-priming grid-enabling effort. This selection will contain at least one code based on each of SPH, static grid and AMR schemes.

The best candidate SPH code for grid-enabling is Matthew Bate's (Exeter). He has recently offered to make a version of his code public as the basis for a common-user SPH code targeted at star and planet formation simulations. The code includes self-gravity, sink particles, and a hierarchical time-stepping scheme to improve performance for problems with a large range in dynamical timescales (for example fragmentation). The code has been widely used on UKAFF to study problems such the collapse of giant molecular clouds, star cluster formation, planet formation, planet-disc interactions, stellar collisions, self-gravitating accretion discs, and stellar winds and outflows. Bate's code shares a common ancestry with other codes used by UKAFF scientists to study white dwarf, neutron star and black hole mergers. This common ancestry will allow all other related codes to be grid-enabled with the minimum of effort.

The representative fixed-grid and AMR codes will be selected after an audit of publicly available codes during the lifetime of the VirtU project. Candidates are the *ZEUS* and *FLASH* codes.

4.7.5 Parallelisation

VirtU will provide access to the wealth of simulation data contained within the TVO, at the simplest level giving access to the raw data in a custom-designed format, but also giving access to refined data products such as group catalogues, merger histories and semi-analytic galaxy populations.

To help the TVO user with their research, we plan to provide a suite of basic software tools which are used commonly by theorists. These include group finders (such as the code used to make merger trees), and codes for evaluating power spectra, correlation functions of objects sets which may be, for example, simulation particles or semi-analytic galaxies.

A number of efficient algorithms for these relatively simple kinds of analyses have been developed over the last few decades. However, most extant codes in use are serial codes. In order for the TVO to be able to handle large datasets it is important to parallelise these codes. We therefore propose in the Project Plan some effort on parallelising the analyses tools. In the case of the group finder, this work is required not only to provide a tool for the user, but also to generate the group catalogues and the merger trees which are fed into the semi-analytic codes in order to generate the prime content of the TVO.

4.8 Visualisation

The importance of versatile analysis and visualization techniques for simulation work is self-evident. Our programme will demand an entirely new level of techniques to cope, for example, with the multi-terabyte datasets generated by the Millennium project, or with the need for visualizing remote datasets.

To give a concrete example, imagine a user wants to analyse an N-body/SPH simulation of the formation of a galaxy for the first time. At the beginning it is highly desirable to obtain an overview of what the dataset as a whole contains - for example, to see what the components of the galaxy look like, where the nearest neighbours of the galaxy are and to watch how the galaxy was assembled. We aim to provide a web-based service for the TVO which will enable interactive exploration of datasets and the making of movies under the directorial control of the user.

To date, most cosmological simulations are particle-based while many fluid dynamic simulations are mesh-based. The visualisation of particle datasets has been somewhat underdeveloped and, while there are plenty of commercial packages for visualizing mesh-based data, to our knowledge, none exist that are suitable for particle-based data. Virgo members have developed state-of-the-art visualization methods that employ adaptively smoothed particle projections to produce high-resolution images and computer animations. These, however, need further development to cope with the Tbyte datasets of the TVO. More sophisticated techniques than those we have used so far for producing 3D smoothed density fields, such as methods based on Wiener filtering or Maximum Entropy, will also need to be developed.

Cosmologists at Durham and Garching, in particular, are known worldwide for the quality of their visualisations, including high-resolution movies of physical processes of DVD or HDTV quality. This material is frequently requested by scientists and the public alike. Last year, Durham cosmologists developed a code for making stereoscopic images which allows the user (with suitable equipment) to view still images or animations from cosmological simulations in 3D. These techniques were used to make a 4-minute stereoscopic movie of the evolution of the Universe which was shown at the Royal Society Summer Science Exhibition last year. Making this movie, however, was a major enterprise which required substantial assistance from a commercial company.

Stereoscopic visualization is valuable not only for spectacular public outreach productions, but also for scientific applications where the ability to visualize complex distributions in 3D can be of great help in developing physical intuition. Auto-stereoscopic 3D displays, now appearing as prototypes in research laboratories and as actual products on the market, bring exciting new opportunities for personal 3D visualisation, where users will be able to sit at their desktop and, unencumbered by glasses or cables, see a true stereoscopic 3D image of their data. However, exploiting this novel hardware requires extensive software development and computer science research. World-leading research experience in the generation of comfortable images for 3D displays is available to the project through applicants from the Durham Computer Science who have pioneered this area. Indeed, there is an ongoing project to develop a grid framework for stereoscopic 3D visualisation for Chemistry applications in the Durham-led "eDemand" project. This project deals with datasets that are orders of magnitude smaller than those that will be typical of VirtU. New strategies and algorithms will need to be developed for coping with such datasets and for mapping the huge depth of regions that characterize cosmological simulations.

Visualization tools are also essential for processing observational datasets and contrasting them with simulations. The galaxy distribution as mapped in the 2dFGRS or SDSS makes up a complex 'cosmic web' of filaments, superclusters and voids. Galaxies of different spectral or morphological types or different luminosities exhibit different clustering patters. Visualizing this web through rotations, zooms and fly-bys is crucial for developing an intuitive feeling for the distribution, suggesting suitable statistics to quantify it and comparing with simulations. Another example of the importance of visualization is provided by the new generation of survey catalogues which will contain hundreds of observed and derived physical parameters for each galaxy. Stereoscopic colour visualization tools, used in conjunction with statistical methods (e.g. Principal Components Analysis) to reduce the number of dimensions, will be very important in analysing the hyper-space of galaxies. Visualization tools are also useful for finding outliers, or new 'strange' objects.

We propose to create a web-based interface and a software pipeline tying together existing image making software to allow the TVO user to produce 2D or 3D images and animations easily from simulation data, the refined data products and observational datasets. This middleware will allow the user to view subsets of real data and compare them directly with mock samples extracted from the TVO. As the size of the simulations increases, it will become impractical for remote users to transfer the data to their local machines via the network. Furthermore, some of the simulation datasets will greatly exceed the memory capacity of a typical workstation. Thus, the visualisation tools should transparently support remote visualisation. The bulk of the computation would be performed on the host supercomputer, and the resulting image stream transferred in real time to the remote workstation.

5 Exploiting VirtU: the theory/observation interface (TOI)

In this section, we describe in more detail how to establish the link between the virtual world of the TVO and the real world of observations. We first develop the concept of the VirtU Telescope Simulator (VTS) and then describe some of the analysis tools that will be made available in VirtU. Many of these can operate equally well on data from the TVO or from the VO. Together, the VTS and analysis tools make up the theory/observation interface or TOI. Like the TVO, the TOI is intended as a dynamic entity that will grow as more and more tools are added.

5.1 The VirtU Telescope Simulator

We propose the development of a tool that will link the Theoretical Virtual Observatory to the observational parts of the Virtual Observatory. The TVO will provide a complete virtual universe with which to compare the real one. This comparison is, however, not trivial since our observed view of the universe is heavily influenced by instrumental and atmospheric effects, particularly from ground-based observations. We propose to develop a set of grid-enabled data-mining tools which will enable realistic simulated data to be generated from the fundamental physical parameters held in the TVO database.

The tools that we plan to develop will be of use to the wider astronomical community, both for the planning of new observations and surveys, using existing instruments and telescopes, and also for the design of new instrumentation and facilities. We forsee these tools being used over the entire lifespan of any astronomical investigation. Firstly, instrument and telescope builders will use them to develop science cases and hence science requirements for future facilities. Then, observers will use them to motivate and justify observing time requests, and finally, both observers and theoreticians will use them to compare the observations and models, and hence develop an improved physical understanding of the universe.

We will also undertake a pilot project to provide grid access to a set of high fidelity simulations of adaptive optics correction on the next generation of extremely large telescopes which will be crucial in optimising the design of these facilities to meet their ambitious scientific goals.

5.1.1 Instrument and Telescope Simulators for 4-8 metre Telescopes

The first phase of this project will be the development of a modular set of simulation tools which will take as input the astrophysical parameters of galaxies from the TVO, and produce as output a series of imaging or spectroscopic observations given a set of instrumental parameters (e.g. aperture size, exposure time, field of view, pixel size, spectral resolution, detector characteristics) chosen by the user. We envisage a generic set of instruments (imagers, high resolution spectrograph, multi-object spectrograph, integral field spectrograph) where the detailed instrument parameters (wavelength range, field of view, pixel size etc.) are controlled by the user via a set of menus. More elaborate instruments or new capabilities would require interaction with the core programming team. The Durham Astronomy Instrumentation Group has considerable expertise in estimating and modelling instrumental effects on both spectroscopy and imaging.

5.1.2 Atmospheric Effects and Adaptive Optics

The modelling of any ground-based optical/near-infrared telescope must include atmospheric effects. Of these effects, spectral modification (absorption and refraction) may be simulated relatively straightforwardly. Atmospheric turbulence, on the other hand, can present a significant challenge, most notably when it is subject to adaptive optics (AO) correction. The image quality correction after AO is necessarily partial and dependent on field position, wavelength, guide-star characteristics, atmospheric parameters and AO system capabilities. The effect of any of these parameters can be, and normally is, decisive in determining the feasibility of any particular observation. Adaptive optics observations is becoming more widely used on existing large 8-m telescopes, and will be crucial for the full exploitation of extremely large telescopes (ELTs) .

The Virtual Telescope Simulator will exploit existing Durham code for 8m atmospheric and AO simulation, which runs on a dedicated cluster. Parallel funding proposals have been made to produce a new hardware-assisted cluster with software capable of detailed ELT AO modelling (scaling to ELT apertures is not possible using conventional codes). The primary requirement is therefore to interface both the existing 8m simulation into the Virtual Telescope and to enable the future ELT simulation to be attached when ready. These interfacing tasks, though greatly simplified compared to the actual AO modelling, are nevertheless non-trivial. To support high fidelity virtual observations, the AO code must be interfaced to the metadata output from the cosmological simulation and must be embedded within the configurable instrumentation model. It will not be sufficient for many design purposes to implement the AO as simply a stage in a pipeline. In some design concepts, AO will be part of the instrument.

A further, and very important consideration will be to provide for future grid-based operation of the virtual-telescope. This will be particularly important when considering the extensive resources which must be deployed for future ELT simulations.

5.2 Distributed application deployment

We now describe a set of tools for analysing and intercomparing real and simulated data in VirtU. This toolkit will grow with time. Here, we describe some examples to be included from the start.

5.2.1 Galaxy SED and image simulation inputs

Numerical simulations are providing increasingly realistic descriptions of the process of star formation and the interaction between stars and their environments. Both of these processes are important for understanding the formation and evolution of galaxies as a whole. Oxford colleagues are building up a library of world-class simulations of processes ranging from the formation of the first stars to the effect of supernovae on the interstellar medium. The scientific returns from this large computational effort will be multiplied manifold by providing public access to the data.

Simply providing raw data is of little use to observers or to theorists interested in comparing observations to current models. Instead, it is much more useful to develop tools for generating 'artificial' observations of the simulated data. This can be done, for example, by providing spatially distributed star formation rates to population synthesis models (see Section 5.2.2) which can then directly generate colours or spectra. The result could be an image of a star forming galaxy at high redshift in colours directly matched to a particular type of observation. This information could then be input into the Virgo simulations of cosmologically large volumes to generate realistic simulated surveys. To carry out this task, we need to not only develop the systematic archive of useful simulations of the TVO, but also generate the tools that can interface with aspects of the AstroGrid. Specific simulations can then be chosen to match astronomical object(s) under observation.

5.2.2 Libraries of synthetic galaxy spectra

Evolutionary population synthesis (EPS) models are used to convert theoretical (whether dynamical or semi-empirical) galaxy evolution models into observable properties, such as brightness and colours. Several popular EPS models exist, e.g. Bruzual-Charlot, Pegase, Padova, Starburst99, Worthy and Yi. They are based on different stellar models and provide different resolution levels and other specific features (e.g. the effects of dust). The EPS codes produce colour and luminosity evolution of hypothetical populations as a function of time and, for each synthetic population, some advanced models (including the one developed by Yi) also generate individual stellar properties, such as mass, temperature, and luminosity. By combining models of galaxy formation and evolution with EPS models, it is possible to infer how the stellar content of galaxies evolves and how it gives rise to the integrated properties which are measured observationally.

We propose to combine different EPS codes and new codes currently under development to improve the spectral resolution so as to match the superb high resolution of modern observed spectra. This combined code would be made available as a VirtU TOI, thereby giving the user the ability to select any desired code or library to mimic the resolution and observational selection effects of the real data (e.g. aperture bias in fibre spectra) and to compress the spectra by extracting traditional diagnostics such as those derived from statistical methods like Principal Component Analysis.

5.2.3 Data compression and classification techniques

With the dramatic growth in multi-wavelength observations of galaxies there is now a huge range of observed and derived properties per object, e.g. luminosity, circular velocity, bulge-to-disk ratio, metallicity, star-formation rate, surface brightness etc. Surveys like 2MASS and SDSS already provide hundreds of parameters per object (of which there are several million in the SDSS), and it is a challenging task to turn this highly dimensional space into a meaningful physical space, in analogy with the H-R diagram for stars. The ultimate goal is to understand the cosmological and astrophysical origin of the physical properties of galaxies and their evolution.

As with data from other fields, there are different approaches for analysing galaxy images and spectra. Conceptually, it is helpful to distinguish between three tasks:

- Data compression
- Classification
- Parameter estimation

The three might of course be related, For example, classification may be performed on a compressed space of spectra, or on the space of astrophysical parameters estimated from the spectra.

Another distinction is between 'unsupervised' methods (in which the data 'speak for themselves', in a model-independent way) and 'supervised' methods (in which training sets of models, or other datasets, are used as a guide). Methods have been developed and utilised to allow 'principled' compression and classification (both supervised and unsupervised) of galaxy spectra and images, such as Principal Component Analysis (PCA), Information Bottleneck and the Fisher information matrix, that could be used to identify physically interesting classes of galaxies. PCA can be generalized to more powerful linear projections, (e.g. projection pursuit) or to nonlinear projections that maximize statistical independence, such as Independent Component Analysis (ICA). These methods provide a low dimensional representation, or compression typically by a factor 100-1000, which greatly facilitates the identification of relevant structure in the dataset. These techniques can also be applied to Integral Field Units, which provide a spectrum per pixel across the galaxy image.

These methods provide powerful tools to mine and analyse the large datasets stored in the Virtual Observatory and Astro-Grid (e.g. 2MASS, SDSS, Vista). They can also identify outliers in a large parameter space, and call attention to 'strange' objects which can be followed up by further observations. Furthermore, they can equally well be used to analyze the simulated data held in the TVO, providing

a direct 'bridge' between models and data. These methods can also be applied to astronomical objects other than galaxies and, in fact, to much wider, non-astronomical datasets.

Using some of the methods described above, the automatic classification of millions of galaxies will become feasible. Other important applications include characterizing the global distribution of galaxies, by deriving, for example, luminosity functions or clustering statistics for samples of galaxies defined according to spectral class or any other physical parameter identified during the classification process.

Algorithms and codes to implement the statistical methods discussed in this section have been written by the applicants. However, they are not optimized for dealing with the huge datasets of the TVO or of observational surveys like SDSS. Extending and adapting them, as well as making them publicly available is a goal of this proposal.

5.2.4 Galaxy parameter estimation

Reliable, physically-based models for galaxy spectra or broad-band photometry will be made available in VirtU (see Section 5.2.1). This resource can be further exploited by providing the requisite tools to estimate key parameters of interest, such as age, metallicity, star-formation rate and dust content, using Maximum Likelihood techniques. This can be done directly from all the spectral bins, or from a linearly compressed version of the data designed to preserve information (in the sense of the 'Fisher Matrix') with respect to the physical parameters of interest.

The sheer size and exquisite quality of modern galaxy data makes it possible, in principle, to estimate an increasing number of physical parameters. As the number of model parameters grows, the problem of optimal parameter estimation in a highly dimensional space becomes increasingly complex from a computational point of view. Fortunately, there are efficient techniques to tackle this problem, such as the Markov Chain Monte Carlo method, which has recently been successfully applied to parameter estimation from the Cosmic Microwave Background. Methods such as this are capable of efficiently generating best-fit parameter values, errorbars and covariance matrices. We plan to make the Markov Chain Monte Carlo, and related parameter estimation methods, available on VirtU. This will require adapting, extending and documenting existing software, as well as developing the appropriate interfaces to the TVO, the VO and the user.

5.2.5 Photometric redshift methods

There is a growing industry based on the estimation of galaxy redshifts from broad-band photometry. The basic photometric redshift technique consists of using the colours of a galaxy in a selection of medium- or broad-band filters as a crude approximation of the galaxy's spectral energy distribution or SED, in order to derive its redshift and spectral type. The technique is very efficient compared to spectroscopy since the signal-to-noise in broad-band filters is much greater than the signal-to-noise in a dispersed spectrum and, furthermore, a whole field of galaxies can be imaged at once while spectroscopy is limited to individual galaxies or those that can be positioned on slits or fibres. However, photometric redshifts are only approximate at best and are sometimes subject to complete misidentifications. For many applications though, large sample sizes are more important than precise redshifts and photometric redshifts may be used to good effect.

Photometric redshifts have been used extensively in recent years on the ultra-deep and well-calibrated Hubble Deep Field observations. The most commonly used approach is the template-fitting technique, implemented, for example in the 'Hyper-z' package. This involves compiling a library of template spectra – either theoretical SEDs from population synthesis models (see Section 5.2.2) or empirical. Then, the expected flux through each survey filter is calculated for each template SED on a grid of redshifts, with corrections for interstellar, intergalactic and Galactic extinction where necessary. A redshift and spectral type are then estimated for each observed galaxy by minimizing χ^2 with respect to redshift, z, and spectral type SED.

Given the growing use of photometric redshifts in giant future surveys such as those to be undertaken by VISTA, it is important to refine existing methods and, if possible, implement better, faster ones. Lahav and collaborators have developed ANN-z, a photo-z method that utilizes Artificial Neural Networks. Unlike the standard template-fitting photometric redshift technique, a large spectroscopically-identified training set is required but, where one is available, ANN-z produces photometric redshift accuracies at least as good as, and often better than the template-fitting method. Furthermore, inputs other than galaxy colours, such as morphology, angular size and surface brightness, may be easily incorporated. The method has been applied successfully to SDSS data as well as to semi-analytic mock galaxy catalogues.

We plan to incorporate into VirtU a package to provide the user with a suite of alternative methods for photo-z estimation, and tools to test them on simulated data. This package will greatly speed up the analysis of large photometric surveys and, at the same time, improve the accuracy of estimated redshifts.

5.2.6 Intergalactic medium and quasar absorption lines

At high redshift, the intergalactic medium (IGM) contains the baryons from which galaxies subsequently form. It is seen in absorption against background quasars, giving rise to a "forest" of Lyman-α and metal absorption lines in their spectra. Structure in the Lyman-α forest has been used to estimate the power spectrum of the dark matter fluctuations which induce it. This, in turn, can be combined with CMB measures of structure on larger scales to estimate fundamental cosmological parameters (such as Ω_0, Λ and the spectral index, n). Thus, quasar absorption lines provide crucial information about galaxy formation and the nature and distribution of dark matter at early times. Their study, a traditional area of extragalactic research, was dramatically transformed when the first realistic N-body/gasdynamic simulations published in the 1990s revealed how the gas is expected to trace the filamentary cosmic web. Since then, observational studies of the IGM have proceeded hand in hand with ever improving simulations. There is a great need for observers and theorists alike to be able to access online the growing library of simulations that can be directly compared with observations. The TVO will satisfy this need.

As part of the TOI, we propose to make available the extensive software that members of Virgo and their colleagues have developed over the years to analyze absorption spectra in the simulations so that they can be directly compared to real data. This includes methods for calculating mock spectra in the simulations at varying resolution, fitting Voigt profiles, estimating neutral gas densities through the pixel-optical-depth method, etc. As discussed in Section 5.1, the VirtU package will allow the user to tailor the results from the simulations to a specific observational setup by allowing them to specify appropriate observational parameters, such as wavelength dependent signal-to-noise, read-out noise, wavelength-coverage including higher-order transitions, continuum-uncertainties, etc. The resulting database of mock spectra and tools to analyze them will be an invaluable resource for IGM research.

6 VO Linkages

In this section, we discuss the connection between VirtU and related activities in the UK and abroad.

6.1 AstroGrid

The Virtual Observatory (VO), represented in the UK by the AstroGrid project, was conceived by the international astronomical community as the means to cope with the flood of data that will be produced by the next generation of astronomical instruments. VirtU aims to provide the theoretical complement to the VO, exploiting the enormous advances in simulation and modelling techniques achieved over the past two decades to produce a unique resource.

Having been conceived, in part, as a complement to AstroGrid, VirtU naturally has close synergies with it. Wherever possible, VirtU will adopt standards and components developed by AstroGrid. In particular, the basic VirtU fabric, addressing data transport and workflow issues, will be taken from the AstroGrid system. Likely AstroGrid components that will be exported include:

- portal architecture
- work flow management
- services for managing distributed user data storage (e.g. MySpace)
- data centre access
- the astropass authorisation system
- the community structure

VirtU will formally interact with the AstroGrid-2 project via a number of mechanisms:

- Membership of the AstroGrid Project Scientist on the VirtU management structure. It is anticipated that AstroGrid will second a percentage of its Project Scientist to the VirtU project, in order to foster the alignment of science priorities for the two projects and facilitate cross-project dialogue
- Membership of a theory-side astronomer on the AstroGrid Science Advisory Group
- Joint meetings between the technical management of AstroGrid-2 and VirtU

6.2 e-Science and grid technology

AstroGrid and the VO are seen as important testbeds and application projects for the development of emerging 'grid' computing technologies (e.g. Hey and Trefethen, 2003). AstroGrid itself is playing a significant role in the generation of new grid technologies. For example, it co-chairs a GGF working group (Work Flow Description Language, WG), is hosting the development work for the Open Grid Services Infrastructure primer, etc. AstroGrid is an early adopter of the Open Grid Services Architecture Data Access and Integration (OGSA-DAI) project deliverables, using these to enable access to data held in distributed databases in a secure grid environment. VirtU will work with AstroGrid in its e-science partnerships, aiming to demonstrate the application of distributed computing techniques in the delivery of theory-side services.

6.2.1 EPCC and Durham Computer Science

VirtU has close connections with the computer science community at both the EPCC and Durham. Virgo has a long-standing, highly successful partnership with the EPCC that has resulted, for example, in the development of a portable MPI version of the *HYDRA* code (www.epcc.ed.ac.uk/computing/research_activities/PPARC/virgo.php). EPCC and Virgo are now involved as partners in the DEISA project, and EPCC leads the OGSA-DAI initiative, all activities that impact on VirtU. Computer science expertise will also be available to VirtU through the Computer Science Department of the University of Durham which, amongst other things, will offer its acknowledged leadership in the development and deployment of stereoscopic visualisation techniques (see Section 4.8)

6.2.2 The DTI/OST/EPSRC e-Science Core Programme - grid network Team

VirtU will input its networking requirements to the UK e-Science Core Programme through its close association with Prof. Peter Clarke at UCL (see http://umbriel.dcs.gla.ac.uk/NeSC/general/teams/gnt).

6.3 DEISA

There are strong links between this proposal and two large European projects. One is the Marie Curie Training Network described below. The other is the Distributed European Infrastructure for Supercomputer Applications (DEISA), a collaboration linking the 8 main supercomputing centres in Europe (CINECA, CNRS-IDRIS, CSC, ECMWF, FZ-Juelich, HPCx, MPG-Garching, and SARA). The objective of this initiative is to provide a persistent and reliable terascale production environment (171 IBM p690 nodes with 5,472 PEs and 7.5 Tbyte) to further the advancement of computational science in Europe. To this effect, a number of Joint Research Activities (JRA) have been set up covering major research

areas, including Cosmology. It is as part of this JRA that EPCC and the Max-Planck Computing Centre at Garching will support the "grid-enablement" of two of the VIRGO consortium's production codes: *GADGET*, and *HYDRA-MPI*. *GADGET* will be ported and optimised for metacomputing environments, whilst *HYDRA-MPI* will be grid-enabled through support for code migration. The DEISA 14 million euro proposal has been submitted to, and is currently under consideration by the European Commission's "Research Infrastructure Action", "Communication Network Development - Grids." Further information about DEISA can be found at www.deisa.org.

While the grid-enablement of *GADGET*, and *HYDRA-MPI* in the context of European supercomputer centres, as defined above, is expected to be fully funded under DEISA, the grid-enabling activities described in Section 4.7 are not included in that programme. Instead, they form part of the present proposal.

6.4 Training Networks

This programme has close links to an application currently under review for a 2 million euro Marie Curie Training Network being proposed by Virgo and collaborators in Europe and the USA for a programme of cosmological supercomputer simulations. Garching and Durham are the two main nodes of this network which seeks support for 8 PhD students and 4 postdocs to work on simulations, including the Millennium project. We anticipate that a number of these students and postdocs will interact with, and provide valuable scientific input to VirtU personnel working at the relevant institutions.

6.5 International Virtual Observatory Alliance

The International Virtual Observatory Alliance (IVOA) has been established by representatives of the major VO projects. Its mission is to "facilitate the international coordination and collaboration required for the development and deployment of the tools, systems and organisational structures necessary to enable the international utilisation of astronomical archives as an integrated interoperating virtual observatory."

VirtU will take an active role in the activities of the IVOA, ensuring the development of standards appropriate for the model data. It is anticipated that VirtU developers will become involved in the appropriate IVOA work groups (see the mailing lists at http://www.ivoa.net/forum/index.html).

6.6 International Partners

This application includes as international partners overseas members of the Virgo consortium and other close associates. They are listed as co-applicants, although, of course, no funding will be directed to them. Their participation in this application is a sign of their commitment to the project and their willingness to contribute to it.

Members of the application team have international partnerships with colleagues involved in VO work in the USA (Djorgovski and Szalay), Canada (Navarro), Mexico (Terlevich), Germany (Steinmetz), Switzerland (Moore, Lilly) and France (Tissier). Through these contacts, some already extending over many years, we will be able to benefit directly from developments in these countries.

7 The longer term future of VirtU

This proposal is for a 3-year programme starting in April 2004. We estimate that this is the length of time required to design, develop and commission an operational TVO and associated TOI. In Section 4.1, we defined a minimum specification for the TVO which we are confident we can deliver in a timely fashion. Throughout this application, we have also indicated future directions in which VirtU could develop.

These involve substantial enhancements in functionality including, for example, the ability to perform new simulations to order or dial-up new semi-analytic catalogues on the grid. The TOI can also be enlarged by the addition of new distributed applications beyond the basic ones that we have outlined in Section 5. These and other exciting extensions will become possible if funding is available beyond the current initiative.

The advent of new observational facilities such as ALMA, ELT, or OWL will demand a new level of theoretical input. It seems certain that as observational resources increase in size and quality so will the hardware required to carry out ever larger simulations. At the same time, it is to be hoped that progress in our understanding of the key physical processes will keep pace with technological developments. Thus, funding permitting, VirtU faces a promising future. As a minimum, VirtU will require continuing support beyond the 3-year period, but if no further development is to take place, a useful structure can be maintained for some time at relatively little cost, probably at the level of 1 or 2 dedicated staff.

8 An inclusive UK initiative: the consortium partners

This application is a collaboration representing most of the UK's astrophysical Cosmology community and their overseas collaborators, a large fraction of the UK's astrophysical gasdynamics community (through UKAFF), and a strong Computer Science component (through the EPCC and the Durham Computer Science department). It has two nuclei: the Virgo consortium for cosmological simulations (of which Frenk is the PI), and a new group that is being established at UCL under the leadership of Lahav which intends to focus on astrophysical modelling. A brief description of the relevant expertise and proposed contribution of each participating institution is given in Appendix A.

9 Project Plan

9.1 Project Management

In the first instance, the overall coordination of the project will be undertaken by Frenk and Lahav who will act as joint Project Leaders. The project will be managed by the Project Management Team consisting of the Project Manager, the Project Scientist (N. Walton) and a Deputy Project Scientist. The later two will be 25% positions. The Project Manager (a post, probably based at Durham, that will be externally advertised) is a 50% position and is most likely to be filled by a senior developer spending the remaining 50% of their time working on the project.

Ultimate responsibility for the delivery of the programme will fall to the VirtU Steering Group (VSG) which will be jointly chaired by Frenk and Lahav and consist of a senior member from each participating institution (e.g. Efstathiou, Jenkins, King, Peacock, Pearce, Silk, Stewart, Thomas and Pringle) or their nominees, the Project Scientist (Walton) and the Deputy Project Scientist. The VSG will invite three external advisors to become members: Professors S. White (Garching, Max Planck Director), A. Lawrence (Edinburgh, AstroGrid Project Leader) and P. Clarke (UCL, GridPP member).

The VSG will act as a Board whose functions will be to:

- define, implement and manage the programme,
- decide on resource allocation,
- organize the recruitment of staff,
- monitor progress.

The VSG will meet as soon as the project is approved to review the Project Plan and to agree on detailed start-up procedures. Regular subsequent meetings at 3-monthly intervals will review progress towards the project's scientific and technical goals, schedule targeted meetings and discuss new initiatives which can further the project's aims and promote dissemination of its results. The VSG will consider whether there is a need to set up a Science Advisory Group made up of people at the collaborating institutions who are

actively involved in the day-to-day activities of VirtU, augmented by other representatives of the user community and international collaborators (such as Drs. V. Springel and A. Evrard). Collaboration-wide meetings will be held every 6 months depending on progress.

9.2 Planning

For the purposes of this proposal, and to aid in its review, the work activities have been broken down into workpackages, as requested by PPARC. However, it is anticipated that the actual management of the programme will follow the procedures that have been successfully developed by AstroGrid, in particular the use of the Unified Process (see http://wiki.astrogrid.org/bin/view/Astrogrid/RbPhaseBPlan for a description). To ensure compatibility, a VirtU representative will sit on the AstroGrid Technical Support Panel (see http://wiki.astrogrid.org/bin/view/AG2/AgConsortiumAndPersonnel). At this stage, it is expected that the functional system will be underpinned by OGSA (the Open Grid Services Architecture) adopted by AstroGrid. Thus, components will be developed with this as a basis.

The main work area strands are then the following:

- Component services: includes construction of the compute grid, data access, job control, etc.
- Component tools: aims to enable existing algorithms and applications in the VirtU environment.
- Advanced data and information visualisation services.
- Research & development: assessment of technologies which are still insufficiently understood to be incorporated directly into the system. This will be an important early activity where the assessment of the appropriateness of uptake of AstroGrid project components for the construction of the base-level TVO will be undertaken. This strand will also include the standards development programme.
- Testbed implementations: involves the implementation of working versions of the software in one or more data centres to test its performance. For instance, VirtU testbed implementations would be tested on an existing "milli-Millennium" simulation (of 10 million particles) at the end of the first year of VirtU.

The following sections describe the main work components that will be delivered by the VirtU project. It is important that each workpackage be read in conjunction with the associated milestones and deliverables as specified in Table 9.3. Note that a major task for the first 6 months will be to define a Road Map for the project which will be developed by the Management team with input from the VSG and the staff in post at the time. Work estimates in the tables are based on an analysis of the work activity, and experience with work carried out within AstroGrid and other e-science programmes.

In the following sections: PM=Project Manager, SRD=Senior researcher/developer, PS=Project Scientist, D=Developer, R=Researcher, D/EU=Developer funded by the EU DEISA project at EPCC. In the tables, **St** denotes the number of full-time staff, working for n **mnth** months, leading to y **St.m** staff months on that topic. In the "notes" entry, the type of staff is assumed to be a software developer or computer scientist unless otherwise stated. For the purposes of project accounting, the project is assumed to begin on 1 April 2004 and have a duration of 3 years, ending March 31, 2007.

9.2.1 VPm: Project Management

Description: This component comprises the project management and project science effort.

Year	Tasks	St	mnth	St.m	Note
1–3	Overall programme management, reporting, etc	0.50	36	18	PM
1–3	Science leadership, SRD, outreach, reporting, interaction with external projects (e.g. EPCC, IVOA) etc	0.50	36	18	PS

9.2.2 VCi: Project coordination infrastructure

Description: The component allows the implementation of project support and control structures.

Year	Tasks	St	mnth	St.m	Note
1–3	Set up project web-pages (e.g. wiki), software repositories, define software and other coding standards. Set up email lists	0.25	36	9	

9.2.3 VAr: VirtU-architecture

Description: This package will develop the VirtU infrastructure, basing it on, and adapted where necessary from, the infrastructure developed within the AstroGrid project.

Year	Tasks	St	mnth	St.m	Note
1	Develop exemplar science tests	1	3	3	R
1–2	Development of VirtU architecture. Based on infrastructure from AstroGrid, adapted where necessary	1	12	12	

9.2.4 VAd: VirtU-architecture Deployment

Description: This defines the VirtU baseline infrastructure, with components largely taken from AstroGrid (end 2004 sees the release of AstroGrid's final iteration product).

Year	Tasks	St	mnth	St.m	Note
1–2	Deployment of AstroGrid components where appropriate: e.g. Job control, myspace, resource registry, user notification etc	2	9	18	
1–2	Authorisation, security model, compatible with AstroPass	1	3	3	

9.2.5 VEi: External Interfaces

Description: This component develops interfaces of TVO and the VO (AstroGrid and Euro-VO). It will focus on areas such as interfacing the TVO and VO registries.

Year	Tasks	St	mnth	St.m	Note
1	Developing interfaces to AstroGrid VO system architecture	1	12	12	

9.2.6 VGr: VirtU-Grid

Description: This component focuses on deploying the VirtU-Grid. A number of key codes will be grid-enabled, and deployed on the testbed beowulf clusters. As a testbed, a number of compute intensive codes will be run on the grid.

Year	Tasks	St	mnth	St.m	Note
2	Construction of VirtU-Grid, linked to VAd, link in VirtU services: processor/disk/algorithms	1	12	12	
1	Grid enabled N-body code (i.e. GADGET)	*1*	*8*	*8*	D/EU
2	Grid enabled hydro code	*1*	*12*	*12*	D/EU
2–3	Deployment across the partners	1	12	12	
2–3	Testbed: enable distributed generation of simulated galaxy spectra over VirtU-grid	1	6	6	

9.2.7 VDa: VirtU-data archive

Description: This component will create an expandable archive containing simulation data. The data will be stored and organised in an efficient way. The VirtU user will need to be able to interrogate the VirtU-registry to find out what simulation data are available and, if authorized, to add new content to the library as well as accessing the data. The TVO will grow as new simulations are added to it. Simulation data are written in many different formats with different conventions for defining the various units. VirtU will create a standard data exchange format for simulation data which will allow seamless exchange of data, compatible with observational standards.

Year	Tasks	St	mnth	St.m	Note
1	Develop archive architecture	2	3	6	
1	Develop efficient formats (e.g. hdf5, binX) for storing Millennium and other simulation data	2	5	10	
1–2	Devise meta-data for standard data model	2	5	10	
2	Create tools to catalogue and archive simulations	2	5	10	
2	Testbed: populate archive with mock galaxy spectra	2	5	10	
2–3	Data annotation	2	3	6	
3	Provide a service to upload new simulation data, including validation	2	7	14	

9.2.8 VPo: VirtU-Portal

Description: The workpackage requires a user interface, the VirtU Portal, be designed and built. The tools and methods for the construction will almost certainly be constrained by the requirement that VirtU should be as compatible as possible with analogous interfaces developed by the AstroGrid team. The interface will allow users to access particular datasets and catalogues and to formulate queries which will return galaxy catalogues or other data products such as images.

Year	Tasks	St	mnth	St.m	Note
1	Initiate design phase of VirtU interface	1	6	6	
1	Liaise with AstroGrid to agree standards, deploy AstroGrid portal components	1	6	6	
2–3	Integration of components e.g. visualisation	1	6	6	
2–3	Build and test the interface, initially using small datasets, then linking to the TVO archive data	1	6	6	

9.2.9 VWo: VirtU-Workflow

Description: VirtU-Workflow - adaptation to grid environment of tools (based on standard techniques) for manipulating simulation data in order to produce usable subsets such as catalogues of groups, merger histories, galaxies, etc. Development of a component management process for complex workflows.

Year	Tasks	St	mnth	St.m	Note
1	Create group catalogues/ merger trees from dark matter simulations	1	6	6	
1	workflow design, investigate astrogrid component	1	4	4	
2	create tools to integrate simulations from different authors	1	4	4	
2	Workflow development	1	4	4	
2	Run semi-analytic code to generate galaxy catalogues	1	4	4	
2	Create galaxy catalogues using simpler algorithms	1	4	4	
3	Workflow deployment	1	4	4	
3	Implement simple analysis tools for dark matter simulations - clustering statistics	1	4	4	

9.2.10 VVs: VirtU-Viz: Stereoscopic visualisation

Description: This component will develop an interface to allow the user to explore 3D datasets interactively and remotely and create animations. Server-side visualisation will enable delivery of images to the end user via a client-side application.

Year	Tasks	St	mnth	St.m	Note
1	Initiate design phase for Visualisation package	1	4	4	
1	Identify a set of capabilities for the visualisation tool	1	2	2	R
2	Develop image and animation engine	1	12	12	
2–3	Build the interface and link into the TVO interface	1	6	6	

9.2.11 VTs: VirtU-Telsim

Description: This component will develop a module to take theoretical data from the TVO and use them to generate 'observed frame' outputs taking account of both instrumental and atmospheric effects.

Year	Tasks	St	mnth	St.m	Note
1	System Design	1	3	3	D/R
1	User interface	1	3	3	D/R
1	Conversion of simulation outputs to data cube	1	6	6	
2	Convolution with telescope optical signatures	1	8	8	
2	Convolution with atmospheric signatures	1	2	2	
3	Convolution with instrumental signatures	1	8	8	
3	Generalised final simulated data product	1	3	3	

9.2.12 VAp: VirtU-Applications

Description: This component will make available key VirtU-enabled applications and tools to allow exploitation and manipulation of theory data held within the TVO.

Year	Tasks	St	mnth	St.m	Note
1	Synthetic galaxy spectra: wrap codes for VirtU deployment	1	4	4	
1	Data Compression: wrap codes (e.g. Wavelet, PCA) for VirtU deployment	1	4	4	
1	Application - redshift methods: wrap codes for VirtU deployment	1	4	4	
2	Provide interface module to allow return of code results to VirtU-portal	1	4	4	
2	Allow instrumental convolution with theory results	2	3	6	
2	Plugin for automated classification - e.g. neural networks	2	3	6	
2	Testbed: data integrator - galaxy type from input parameters	2	3	6	
3	Redshift methods: enable result return from redshift app applied to real data (AstroGrid) & simulated (VirtU)	2	4	8	
3	Provide tools to extract spectral diagnostics (e.g. PCA)	2	4	8	
3	Interface module to allow integration of synthetic spectra into mock halos selected from N-body simulations	2	3	6	

9.2.13 VRo: VirtU-Robust

Description: This component addresses robustness of the system. Areas of work include maintenance, fault tolerance and data duplication, documentation.

Year	Tasks	St	mnth	St.m	Note
1–3	VirtU system documentation	1	4	4	
2	Fault tolerance, data duplication, security	1	3	3	
1–3	VirtU system maintainance, development of quality assurance process	1	3	3	

9.3 Milestones & Deliverable

From the breakdown of work components it is possible to give an initial estimate of milestones and key deliverables. Key metrics to enable the measurement of progress against milestones will be developed and published in the early stages of the project planning process. The project will run as a series of three monthly iterations with the first completing end June 2004.

Date	Comp.	Milestone	Deliverable
May-2004	VCi	Project web presence (e.g. wiki)	Interactive tools available
Jun-2004	All	Iteration 1:	Project staff in place
Jul-2004	VAr	Science Requirements	SRD available
Sep-2004	VPm	Full project roadmap	Iteration deliverables published
Oct-2004	VWo	Millennium sim fully processed	Minimum TVO content delivered
Jan-2005	VVs	Stereoscopic visualisation design frozen	Design document
Mar-2005	VAr	VirtU architecture defined	Architecture described in UML
Mar-2005	VAp	VirtU spectral libraries prototype	Resource
Apr-2005	VRo	Audit of system reliability	Security strategy document
Jun-2005	VAd	Initial architecture components deployed	Limited functionality, eg astropass, some VirtU-space available
Jul-2006	VGr	First grid enabled N-body codes (DEISA)	Code demonstration
Sep-2005	All	Iteration 6: VirtU: Release 0.1	First system deployed
Dec-2005	VPo	Iteration 7: Visualisation	First server-based capability
Dec-2005	VTs	First light VirtU-Telsim	First realistic images from TVO
Jan-2006	VDa	Testbed: QSO mock spectra archive	Resource available for interrogation
Mar-2006	All	Iteration 8: VirtU: Release 0.2	Limited functionality available
Mar-2006	VAp	VirtU diagnostic tools	PCA available
Mar-2006	VAp	Limited compression codes deployed	Prototype available
Aug-2006	All	VirtU: Release 0.5	Science grade system available
Dec-2006	All	Draft Documentation	System Documentation
Dec-2006	VAp	Service implementing distance calculator	feature integrated in VirtU
Jan-2007	All	Iteration 11:	VirtU version 1.0 deployed
Mar-2007	All	VirtU system completed	System available to end users
Apr-2007	All	Iteration 12:	VirtU version 1.0 operational, docs

9.4 Risk Analysis

A full risk analysis will be developed in the initial planning phase of VirtU. At this stage a top-level risk matrix is indicated; some risks and impact are similar to those noted by the AstroGrid project.

Risk	**Prob**	**Impact**	**Notes**
Technical			
OGSA is late in delivering	med	med	Switch to web services
Large simulation data is delayed (e.g. the Millennium run)	Low	Med	Smaller simulations will be used as testbeds.
Operational			
SRIF-2 award to ICC withdrawn for lack of 10% PPARC contribution	Low	High	Find alternative funding for TVO hardware
Social			
Divergence with European partners, e.g. GAVO	Low	Med	Coordination through e.g. IVOA will minimise risk
Personnel			
Difficulty in hiring staff	Low	High	Current IT situation is such that the hiring situation is good
Activity Dependencies			
AstroGrid is late in delivering	Low	High	Develop own min architecture, de-scope
AstroGrid2 is not funded	Low	High	Re-scope during planning phase
Budget			
Project is late	Low	High	Iterative project ends at a given date, possibly with descoped deliverables
Project overbudget	Low	High	Management/budget control in place
H/W, S/W costs increase	Med	Low	These costs small fraction of budget

9.5 Personnel Request

The above component effort analysis determines the personnel resource request as follows. The salary scales are in line with those in operation in the AstroGrid project. The salaries are below industry norms, but are sufficient to attract well qualified software engineers.

Type	Grade	Pt	Salary (8/2002)	Overhead (@46%)	NIC (@21.5%)	Add Pts (yrs 2+3)	Pay Inflation @3%	Total (3yrs) K£
PM/SD	RA3	21	37	17	8	5	4	195
PS	RA3	21	18.5 (50%)	8.5	4	2.5	2	97.5
SD	RA2	16	30.5	14	6.5	4	4	161
R	RA1A	11	25.5	11.5	5.5	3	3	133.5
CS	RA1A	11	25.5	11.5	5.5	3	3	133.5
D	RA1A	11	25.5	11.5	5.5	3	3	133.5
D	RA1A	11	25.5	11.5	5.5	3	3	133.5
D	RA1A	11	25.5	11.5	5.5	3	3	133.5
D	RA1A	11	25.5	11.5	5.5	3	3	133.5
D	RA1A	11	25.5	11.5	5.5	3	3	133.5
D	RA1A	11	25.5	11.5	5.5	3	3	133.5
	Support staff rate (10% sysadmin, 8% secr.) $10 \times 3 \times$ £8 000:							240
							Total:	1761.5

D in the table above is a software developer, CS a computer scientist, and R an astronomer researcher with computer background. SD is the senior developer, PS the project scientist (split with Walton as 25% FTE PS, at Pt 23, and TBD as the 25% FTE deputy PS, at Pt 17), and PM/SD the project manager/senior developer role. Walton's role will be, in part, to ensure scientific synergy between AstroGrid and VirtU at the observational/theory VO interface.

It is anticipated that the core project management team will be split across Durham, UCL and Cambridge. Durham and UCL will host the core developer teams, with smaller deployments at other consortium partners (e.g. Cambridge, UKAFF, EPCC, etc). This deployment model is based upon that which has been successfully implemented in the AstroGrid project. Final resource decisions will be made by the VirtU Steering Group at the commencement of the project.

9.6 Equipment And Travel

9.6.1 Equipment: justification and request

The primary hardware requirement for VirtU is a storage system capable of handling and providing efficient access to the TVO data. To keep the costs down, we propose a hybrid system, with a fast access 20 Tbyte primary component and a slow access secondary component, also of size 20Tbyte. The primary component will be used to store frequently accessed data requiring manipulation in, for example, computationally-intensive queries while the secondary component will be used to store less frequently accessed data such as the raw output from the Millennium simulation. Data will be backed-up in a tape system which is significantly cheaper than even the slow disk. The tapes will store one-off backups of static raw data like the Millennium simulation as well as more volatile and compute-expensive data. We also propose to purchase 4 slow RAID disk systems to be dispersed in VirtU sites to enable subsets of the data to be held locally for experimentation with distributed computing. To save costs, purchase of the primary storage system will be staged during years 1 and 2 to match progress in the construction of the TVO software infrastructure, as detailed in Section 9.3.

The primary disk storage will be installed in the machine room at the e-Science Research Institute at Durham which will also house the main ICC supercomputers. It will be served by a small, 4-processor server. This system will be able cope with modest demands, but will be insufficient to deal with queries requiring substantial computation. To keep the hardware costs of this proposal to a minimum, the ICC is prepared to divert some of its supercomputing resources to TVO use. In particular, part of the supercomputer to be purchased with the £1 Million SRIF-2 provisional award offered to the ICC

by Durham University (see Appendix A.1), will be made available to this project. We stress that, in accordance with HEFCE's rules, the SRIF-2 award to the ICC is conditional on the ICC finding 10% of the funds from external sources. The request made here will provide the 10% contribution; *without it there is a real risk that the University will withdraw its SRIF-2 award.*

We are also requesting 2 small beowulf systems to be sited at UCL and Sussex or Edinburgh in order to provide small-scale versions of the Durham machine for development activities by locally based project staff, including distributed applications. All prices below include VAT and 3-year warranty.

A. The TVO server, data storage and backup (Durham)

– 20 Tbyte primary data storage. IA64 server with 4 processors and 16Gbytes of RAM connecting, through 4 fibre channels via a fibre channel switch, to 10 2Tbyte RAID boxes, delivering a maximum rate of 1.6 Gbytes per second. (£147,800).

– Software to manage primary data storage: Veritas volume manager, server & file system (£15,500)

– 20 Tbytes of slow RAID disk storage (£48,000).

– 15 slot, single drive expandable tape library (£15,000)

– 30 Tbyte Sony S-AIT high capacity tapes (£12,000)

B. Local development systems (UCL, Sussex or Edinburgh)

– 16-processor Beowulf cluster connected with Gbit network (two off £13,500)

C. Local VirtU-Space storage (Cambridge, Nottingham, UCL, Oxford)

– 3.2 Tbyte of slow RAID disk storage (4 off £8,000)

D. Staff equipment costs

Staff	PC/Laptop (K£)	Consumables (K£)	Inst. Contribution (K£)	Staff No.	Total (K£)
VirtU PM/SD+PS+DPS	2.5	1.0 pa	1.0 pa	1+0.25+0.25	12.8
Developers	1.75	1.0 pa	1.0 pa	8	62.0
				Total:	**74.8**

9.6.2 Travel

Staff	UK Travel (K£)	Int'l Travel (K£)	Staff No.	Total (K£)
VirtU PM/SD+PS+DPS	3.0 pa	4.0 pa	1+0.25+0.25	32
Developers	2.0 pa	1.0 pa	8	72
Lead Investigators	1.0 pa	1.5 pa	8	60
			Total:	**164**

9.7 Budget Request

The total budget request for the project of £**2.3M** is calculated from the numbers given in the previous sections. We assume a three year project commencing 1 April 2004 and ending 31 March 2007.

Item	Cost (K£)
Staff	1761.5
Travel	164.0
Equipment (staff)	74.8
Equipment (compute)	297.3
Total:	**2297.6**

10 References

AstroGrid - http://www.astrogrid.org

Astrophysical Virtual Observatory (AVO) - http://www.euro-vo.org

DEISA – http://www.deisa.org/

Euro-VO - http://www.euro-vo.org

Frenk, C.S., Colberg, J.M., Couchman, H.M.P., Efstathiou, G., Evrard, A.E., Jenkins, A., MacFarland, T., Moore, B., Peacock, J., Pearce, F.R., Thomas, P.A., White, S.D.M. & Yoshida, N. (2000), *Public Release of N-body simulation and related data by the Virgo consortium*, astro-ph/0007362

Frenk, C.S., White, S.D.M., Bode, P., Bond, R.J., Bryan, G.L., Cen, R., Couchman, H.M.P., Evrard, A.E., Gnedin, N. Jenkins, A., Khokhlov, A.M., Klypin, A., Navarro, J.F., Norman, M.L., Ostriker, J.P., Owen, J.M., Pearce, F.R., Pen, U.-L., Steinmetz, M., Thomas, P.A., Villumsen, J.V., Wadsley, J.W., Warren, M.S., Xu, G. & Yepes, G. (1999), *The Santa Barbara cluster comparison project: a test of cosmological hydrodynamics codes*, Astrophysical Journal, 525, 554-582.

Global Grid Forum (GGF) - http://www.gridforum.org

Globus - http://www.globus.org

Hey, A., Trefethen, A., 2003, in 'Grid Computing, Making the Global Infrastructure a Reality', eds Fran Berman, Geoffrey Fox, Anthony Hey, Wiley, p809.

International Virtual Observatory Alliance (IVOA) - http://www.ivoa.net

National Virtual Observatory (NVO) - http://www.us-vo.org

Open Grid Services Architecture Data Access and Integration (OGSA-DAI) - see http://www.ogsa-dai.org.uk/

VOTable - see http://cdsweb.u-strasbg.fr/doc/VOTable/

Williams, R, 2003, in 'Grid Computing, Making the Global Infrastructure a Reality', eds Fran Berman, Geoffrey Fox, Anthony Hey, Wiley, p837

Appendix A: The consortium partners

A.1 Durham

Durham's contribution to this proposal comes from both the Physics and Computer Science departments. Durham is the UK base of the Virgo consortium. The recently founded Institute for Computational Cosmology (of which Frenk is Director) hosts the largest group in the UK, also one of the largest in the world, working on cosmological supercomputers simulations. The ICC is part of the Ogden Centre for Fundamental Physics housed in a new building.

The University of Durham has identified e-science as a high priority area and invested over £2 million of SRIF-1 funds on another new building to house the Durham e-Science Research Institute which will be inaugurated this Summer. e-Science Professor, I. Stewart, a co-applicant of this proposal, will be the first Director of the Institute. Durham Computer scientists are already involved in major e-science programmes. For example, Prof. Jie Xu leads the EPSRC/DTI e-Demand project, funded by the UK e-science Core Programme, to build a service-based architecture for dependable e-science applications and he is co-leader of the EPSRC IBHIS project on large-scale information integration in a heterogeneous database environment. He is the Durham representative on the Executive Board of the EPSRC North East e-Science Centre.

It is expected that some of the staff funded under this proposal will be sited in the new multidisciplinary Durham e-Science Center where they will be able to interact with colleagues from computer science and other subjects. We also expect that the University will create new lectureships associated with e-science, possibly including, if this proposal succeeds, one associated with the VirtU programme.

The ICC has been provisionally awarded £1 million of SRIF-2 money to purchase a replacement for the Cosmology Machine which will become obsolete towards the end of the SRIF-2 period. Suitably extended, this machine will become the hub of the TVO. However, the SRIF-2 award is contingent on the ICC securing 10% of the costs from external sources. Part of the hardware requested in this application is intended specifically to add the TVO functionality to the new machine. This hardware would fulfill the 10% requirement for the SRIF-2 support, thus providing a large multiplicative factor for a relatively modest investment by PPARC.

A.2 UCL

University College London is well placed to coordinate and plan the Theory/Observations Interface (TOI) part of VirtU. Lahav (currently at IoA Cambridge) has just been appointed as the Perren Professor of Astronomy at UCL, from 1 October 2003. e-Science in general, and VirtU in particular, are key projects for the new group at UCL. Lahav is a team member of the 2dF galaxy redshift survey, aspects of which can be viewed as proto-types for the TOI. Other groups in UCL, in particular the particle physics group and MSSL, are also active in e-science programmes (GridPP and AstroGrid).

A.3 Cambridge

The Institute of Astronomy in Cambridge is one of the largest European centres in the field of Cosmology and galaxy formation. For this proposal, state-of-the-art N-body and hydrodynamic simulations will be supplied for the TVO by G. Efstathiou, J. Ostriker and collaborators. Cambridge is also the base of N. Walton, the AstroGrid Project Scientist. The IoA have close links with the Cambridge e-Science Centre, Walton being a member of its Board of Management. The IoA also have links with the Computer Lab, and Microsoft Research Centre. The IoA is host to one of the UK's premier data centres, being the primary point of access to many of the UK's major optical and infrared datasets (e.g. from ING, UKIRT, AAO, etc).

A.4 Leicester

The Department of Physics and Astronomy at the University of Leicester houses UKAFF and is also a major partner in the AstroGrid project. This, and the involvement in astronomical database/archiving

with projects such as XMM, Swift & LEDAS, has resulted in the Department becoming one of the new e-Science Centres of Excellence. The location of UKAFF at Leicester builds further on the strength of the Theoretical Astrophysics Group led by Prof. Andrew King and their expertise in modelling astrophysical fluids. The University is also committing £1.2M of its SRIF-2 allocation to high performance computing within the Physics and Astronomy Department.

A.5 Edinburgh and EPCC

Edinburgh is represented by J. Peacock from the Department of Physics and Astronomy, and J.C. Desplat and G. Pringle from the Edinburgh Parallel Computing Centre. Edinburgh is also where A. Lawrence, the AstroGrid Project Leader, is based and is host to a substantial AstroGrid development group. Peacock is the UK PI of the 2dFGRS and a member of Virgo and Desplat and Pringle are involved in DEISA.

EPCC is one of the five funding partners of the UK National e-Science Centre (NeSC) and, as such, has taken an active role in stimulating and sustaining the development of e-Science in the UK, contributing significantly to its international profile and ensuring that its techniques are rapidly propagated to commerce and industry. EPCC staff are actively involved in developing open-source Grid middleware – perhaps most notably OGSA-DAI which is distributed with the Globus toolkit – and in establishing standards for future software development via active participation in the Global Grid Forum (GGF).

EPCC is engaged in a range of activities of direct relevance to VirtU. In addition to their planned participation in DEISA (see Section 6.3), EPCC is spearheading an initiative, just officially approved by the GGF Steeting Committee, to form a new GGF working group to define the Data Format Description Language (DFDL; see http://www.epcc.ed.ac.uk/dfdl/), a generic XML-based language for describing the structure of binary and character encoded files and data streams. This initiative was prompted by successful attempts to demonstrate the benefits of using the Binary XML Description Language (BinX), also developed at EPCC, for astronomical data (see http://www.edikt.org/binx/introduction.htm).

Finally, EPCC is involved in the GridPP consortium through its QCD Grid project. This project is particularly relevant to VirtU as it aims to deploy a secure, reliable and expandable multi-terabyte distributed-storage resource atop which lies an XML Database Server (XDS) for storing and querying the metadata, along with a set of graphical and command-line tools by which researchers may store, query and retrieve the data held on the Grid (see http://www.gridpp.ac.uk/qcdgrid/).

A.6 Sussex

Sussex has extensive experience in the development of parallel N-body/hydrodynamics codes and in performing and analyzing large simulations. This experience will prove invaluable in the design and production of data-analysis pipelines for hydrodynamics data, most notably the generation of X-ray and Sunyaev-Zel'dovich catalogues. Sussex is committed to e-science and has a good record of putting JIF resources into this area.

A.7 Oxford

Oxford Astrophysics is the focus of the UK's strategic investment in the FMOS project for the Subaru telescope, and has strong links to the UK Gemini programme (through the UK Gemini Support Group), and the VISTA programme (through the joint appointment of the FMOS P.I., G. Dalton, to the VISTA IR Camera Scientist post at RAL). They are using these programmes to build on their long standing experience with large-scale survey programmes (APM, 2dFGRS, 2dFQSO, INT Surveys, UKIDSS), with the logical next step being their ongoing development of next-generation survey instrumentation for the VLT (MUSE, KMOS). Oxford also has a large simulation group that specialises in hydrodynamic simulations of star formation on all scales, and unequalled UK expertise in stellar population synthesis. These groups will provide the simulations and SED input needed for the TOI.

A.8 Nottingham

Nottingham has just announced a major e-science initiative, with 2.5 million of SRIF-2 money committed to it. The focus of the Nottingham proposal is on remote sensing, visualisation and data manipulation, the latter two areas of key importance to this project. Half of the funds are earmarked for the purchase of terascale computing resources.