STFC HPC Requirements looking towards future provision of HPC facilities

Prepared by the STFC Computing Advisory Panel (CAP)
April 11 2011

Status: Final Version

1. Introduction

During 2011 the Research Councils will be involved in planning discussions regarding future High Performance Computing (HPC) provision on both a National and European Scale. STFC supports several areas of high profile science which depend critically upon HPC facilities. It is therefore essential that STFC has available a summary of the HPC requirements which follow from the science challenges across STFC, and advice from the community to assist in such discussions.

The STFC Computing Advisory Panel (CAP) has been asked to provide this information. CAP is aware of many documents which have reviewed STFC HPC science, and of the latest strategy document from 2006. We have referenced these and the aim of this document is not to repeat the information contained therein.

This document only addresses the HPC computing requirement and explicitly does not address the science cases or the prioritisation thereof.

Members of CAP are listed in Appendix 1.

2. Existing documents

Several reviews have taken place concerning STFC (and previously PPARC) science which uses HPC computing:

PPARC High Performance Computing Review, B Webber, EWN Glover, A Liddle, E Priest, April 2003: http://www.pparc.ac.uk/Rs/Cm/ReptR/HPCRep.pdf

The Strategic Framework for High End Computing (HEC), High End Computing Strategy Committee, 2006 mhttp://www.epsrc.ac.uk/ResearchFunding/FacilitiesAndServices/HighPerformanceComputing/HPCStrategy/2006StrategicFramework.htm

International Review of Research Using HPC in the UK, EPSRC and DFG, December 2005 http://www.epsrc.ac.uk/ResearchFunding/FacilitiesAndServices/HighPerformanceComputing/InternationalReview/default.htm

Supporting a Competitive PPARC Theory Programme: A Strategy for High Performance Computing, report of the HPC Working Group, 2006
www.pparc.ac.uk/rs/fs/scptprep.pdf

The Executive Summary from the last of these is given in Appendix 2.

3. Science Challenges

As stated above we do not address the science cases or prioritisation thereof. We give only a very brief summary to set the context.

- The study of cosmology and astronomy where challenges include: modelling the large-scale structure of the Universe; the formation and evolution of galaxies and clusters of galaxies, and the physics of the intergalactic and intracluster gas; studying fundamental physics in the early universe; the effects of dark energy and modified gravity on structure formation; parameter estimation and model selection using large astronomical datasets generated by cosmic microwave background experiments and large galaxy redshift surveys; and theoretical simulations vital for efficient and effective use of STFC supported telescope facilities on ground or in space.

- The study of the 'local' astronomical environment – from the upper levels of the Earth's atmosphere to the stars and gas clouds of the galaxy. Modelling solar and planetary magnetohydrodynamics and galactic dynamos; modelling star and planet formation; understanding the local astronomical environment from the upper levels of the Earth's atmosphere, to space weather, to gas and dust around stars.

- The study of the theory of the strong force using the numerical techniques of Lattice QCD. The objective is to increase the predictive power of the Standard Model of particle physics and other relevant strongly-interacting field theories by numerical simulation of lattice-regularised quantum field theory. This leads to the calculation of physical quantities which are essential to the interpretation of experimental observations. The basic numerical problem is the estimation of a huge multi-dimensional integral over quark and gluon degrees of freedom using Monte Carlo importance sampling (the dimension of the integral is proportional to the volume of spacetime under consideration). The most computationally-demanding step is the repeated inversion of a large sparse matrix describing quark propagation through a background gluon field configuration.

- Support for the STFC National facilities including Diamond, ISIS and the Central Laser Facility (CLF). These facilities provide the large-scale experimental instrumentation needed to underpin a wide range of research activities in a number of key areas of science and technology, including physics, chemistry, biology and earth sciences. General themes include studies of the structure and dynamics of condensed matter, molecular and plasma systems using high intensity light and neutron beamlines.

4. The nature of the HPC computing requirement

4.1. Architectures

The details of the technical computing challenge and the scales and architectures of computing required in each science sector are given in Appendix 3.

Different types of task are suited to different architectures depending upon the compute task. Before we comment upon these we feel it is important to point out that the choice of architectures is not completely separable from the prevailing funding model. Given a particular set of constraints then the choice which provides most cost effectiveness *as seen by the research grant* may vary.

- A large part of the STFC HPC task is well suited to commodity processors running Linux, with high speed – low latency inter-processor communications. The "commodity processor" requirement is driven by a combination of price performance and the performance characteristics of the production codes which are parallelised using MPI. The high speed – low latency inter-processor communications are today satisfied by high-speed backplane interconnects (today this means Myrinet or Infiniband). Currently Infiniband is particularly attractive because it provides significantly better performance for both bandwidth and latency and at the same time good price performance (the use of slower interconnects such as gigabit Ethernet impacts significantly on the scaling of the production codes to many cores, both because of the limited bandwidth and for many applications the relatively large latency). Overall this kind of architecture is suited to (i) much of the cosmology work (ii) the magneto-hydrodynamics work (iii) Lattice QCD analysis and for the exploration of new lattice action formulations and new physical problems, where speed and ease of code development is the important (iv) the STFC facilities requirements. Most often this type of facility is realised today on large "commodity clusters", i.e. large clusters of inexpensive servers.

- Optimised design machines are used particularly within the Lattice QCD sector. The main "production simulation" is ideally suited to Single Instruction Multiple Data (SIMD) parallel computation on many nodes; the essential requirements are good memory bandwidth on each node (around 1 B/s per sustained flop/s) with fast inter-processor communication (around 1 B/s per 10 sustained flop/s). Communication is mainly nearest-neighbour with occasional global sums. With enough computational nodes, the memory footprint on each node is small and a large cache can be highly beneficial. In the case of UKQCD, the most computationally-intensive simulations have until recently mainly been carried out on a dedicated specialised machine (QCDOC), and have now been switched to machines based on similar design principles but supplied by a major commercial vendor. These machines are very cost effective, with electricity consumption roughly 10%-30% of that of a PC cluster with equivalent computing power.

- The complementary architecture of large shared memory machines (SMP) continues to have an important role. Such systems support all parallel programming approaches equally well, can simplify code development. As a concrete example, an SMP architecture greatly simplified one critical step in the evaluation of the WMAP CMB trispectrum by COSMOS, allowing each core rapid access to a huge dataset. Alternate approaches involving either extensive I/O or coding a complete domain decomposition, would have delayed the science result preventing inclusion in the first Planck cosmology papers.

- It is important to note that many of the science sectors operate by generating a large simulated data set once in a production, and then analysing this many times to produce many research results. This means there is a requirement for post-generation storage facilities, and computing facilities, which can be just as computationally intensive as the generation stage. Both HPC and non-HPC machines are used here.

- Although a relatively new area, the use of General Purpose Graphics Processing Units (GPGPUs) is set to play an important role for simulation and analysis. These potentially provide very high cost effectiveness and many science sectors are investing effort to develop their use their use. This fact re-enforces the need to maintain agility in strategy (i.e. to be able to react to opportunities to exploit step changes in cost-effectiveness).

4.2. Scale & Agility

As it typical of very many compute intensive science areas, the nature of the task requires access to different scales of resource at different times during the lifecycle of the research. By and large the requirement ranges from large scale/lower agility --> small scale/high agility.

- At the early stages, typically development stages, the requirement is for relatively small resources, but with frequent and easy access, and with fast turn around (i.e. small scale/large agility). This is vital for efficient code testing and debugging which requires a turn around within a day, if not hours. Another use case is where a large number of specific code libraries are needed which may not be present on centralized machines.

- There is often a middle stage after development, e.g. parameter space exploration, or validation and testing of pre-production. Naturally enough this requires access to mid range resources at modest frequency. There may also be smaller scale productions suited to this (i.e. medium scale/medium agility) where some 10s to 100s of cores for a several weeks are needed.

- Finally, a task may move into production, by which we mean major simulations to produce data upon which subsequent analysis is carried out. As example this includes the generation of a large ensemble of gauge field configurations in LQCD, or the generation of a long time sequence of matter distribution for cosmology. This stage is

suited to large facilities and does not have an immediate access requirement in general (i.e. large scale/small agility)

The characterisation above is of course not exhaustive, but serves to exemplify the range of access patterns. The overall message which has been stated many times in the past, and which has been affirmed with the community, is that a range of scales of machine is needed. If would be false to assume that collecting everything together in one place would lead to efficiency in terms of either cost or research competitiveness

4.3. Other Issues

The following other important issues are noted:

- It is essential that STFC embraces the need to develop a consistent and holistic funding model which deals with both the non-recurrent and recurrent costs of computing.

- The competition for access to international facilities is very intense and, in addition to excellent and ambitious science, will increasingly favour projects that have well organised software and support (at local and/or national levels).

- Exascale computing brings the need for radical changes to algorithms and software to address the need for fault tolerance, strong scaling and much steeper memory hierarchies (the latter two do not affect just the high-end machines, since the core count on commodity processors will grow rapidly).

5. Future HPC Resources

In this section we consider how different putative future facilities could satisfy the requirements, and what broad capabilities would be needed to deliver the science.

5.1. International HPC facilities (e.g. PRACE)

The UK is a member of PRACE. As such UK researchers now have access to different types of supercomputing architectures, managed through a central peer review system. Four of the 19 members have agreed to contribute supercomputers (France, Germany, Italy and Spain) and are termed hosting partners, and in addition up to 70M Euro of funding from the European Commission is supporting the implementation of the infrastructure.

Such a large scale International facility could almost certainly technically provide the resource for running a large fraction of the HPC workload, including “high profile” productions in LQCD, cosmology, particle cosmology, galactic structure and hydrodynamics.

The LQCD community has already obtained (competitive) allocations from DEISA on international resources, which proved to be very useful. Much of the LQCD sector collaborates

with international partners who pool resources already. There is, therefore, no reason to imagine that PRACE could not provide a valuable addition in this sector.

The Virgo Consortium has made use of a number of international scale HPC facilities for Grand Challenge Projects. The include the GIF project at EPCC, the Hubble volume and Millennium simulations at the RZG Garching, STELLA (the LOFAR correlator) and the LRZ Garching for the Aquarius project, HPCx for the GIMIC project (through DEISA) and most recently the MXXL simulation on Juropa at Juelich. For these kinds of project access to the HPC facility is typically negotiated over some agreed time, and the simulation data generated during the project shipped back to local systems such as COSMA at Durham for post-processing and data curation.

Such a facility would be unlikely to satisfy the small scale-agile access requirement unless a novel partitioning of scheduling classes were provided, e.g. a special class to make it appear as if a more local virtual non-shared machine were available at all times.

It would also not be suitable for the STFC Facilities access pattern unless a novel partitioning of access scheduling classes were provided to give neo-real time access during instrument operation. Even then the inability to co-schedule maintenance and upgrade times with instrument runs would be a potential problem. There would also be unnecessary network dependence.

A suitable cost-at-point-of-use-model would have to be available.

If this was the only facility available in the future then we make the following comments concerning the likely impact upon science delivery:

- It could provide a valuable resource suited to a large fraction of production work.
- There would be a potential loss of cost effectiveness as access to the facility may cost more per unit resource than is needed for many applications, or for the development phase.
- Given that access to such facilities is generally through ongoing peer review, it is essential that these review processes acknowledge the role of the facility in providing agile and responsive development and exploration facilities. This is connected with the previous point as such work may not appear, at first sight to optimally exploit the large scale systems, but is essential for that exploitation.
- As a part of this it would be necessary to allow flexibility in the terms of use post award (e.g. an unforeseen line of investigation emerges, or for some competitive reason) without going back to the peer review.

5.2. National HPC facilities

The assessment in this section is very similar to that of the International Facility and we do not repeat those points here.

Two historical points which indicate barriers are

- Cost of using the national facility is most often cited as a problem. Commodity clusters appear to give better performance/cost ratio as seen by the grant holder. Another example comes from LQCD where the cost of say 100 Tflop years of sustained computing effort on a lattice QCD project on the IBM BlueGene/Q system to be installed by UKQCD is estimated to be at least an order of magnitude cheaper than HECToR access costs under the prevailing model.
- To date HECToR has not been used by the Virgo consortium, but could in principle have been used for several recent Grand Challenge projects. As currently configured, HECToR does not provide sufficient disk/archival space to cope with all of the simulation data Virgo has produced and lacks the infrastructure for some post-processing tasks which require large shared memory machines for reasons of agility.

If this was the only facility available in the future then we make the following comments concerning the likely impact upon science delivery:

- It could provide a valuable resource suited to a large fraction of production work.
- There would be a potential loss of cost effectiveness as access to the facility may cost more per unit resource than is needed for many applications, or for the development phase. Much would depend upon the mechanisms for awarding time and the cost model.
- Mechanisms would need to be enacted to ensure quick turnaround access during development.
- It is not axiomatic that such a facility need be a single machine. Consideration should be given to a complementary set of machines to encompass the entire science lifecycle.

5.3. Consortia based HPC computing facilities i.e. shared HPC across one or more University dedicated to a broad science sector

Consortia based facilities have already been demonstrated to provide for the requirements of the different science sectors up to now (by design). These consortia facilities tend to be large systems shared across several institutes dedicated to a broad science sector. These include COSMOS, HORIZON, MIRACLE, UKMHD, UKQCD and VIRGO. These consortia, along with other groups, now collaborate through the DiRAC project.

The majority of these science areas have been well served in the past by Commodity Clusters with thousands of cores, with (today) Myrinet or Infiniband backplanes which provide adequate communications speeds. Such clusters have proven to be both architecturally suited, and to provide the most cost effective solution within the current funding model. There is significant competition in the market and many leading institutions procure such facilities. These have provided well for the agile access requirement. They have also provided a valuable focus for expertise and training.

In the case of UKQCD the consortium facilities have been essential to provide the necessary resources. Even so, until the advent of the DiRAC project, the total number of flops available

had fallen below that available to international competitors, and the consortium had been increasingly reliant upon international collaborators since 2004/5 for production runs.

5.4. Other forms of computing e.g. Grid or cloud

The Grid (e.g. the Worldwide LHC Grid – WLCG)) is well suited to massively parallel work with no inter-processor communication. Whilst the number of nodes on the WLCG is very large (~200,000) it is generally not possible to co-schedule nodes today, and even if one could the number which could be co-scheduled in close proximity is small. Inter processor communications would compete with other traffic. This is therefore unsuitable for STFC HPC Science.

Cloud computing undoubtedly has a place in the future. However today it is far too underdeveloped in terms of an access model for large scale science computing. There is currently no viable model for high performance HPC cloud computing with guarantees on inter-processor communications. We believe it is important for STFC to engage, in concert with the RCs and possibly HEFCE, to understand how cloud computing can contribute to reducing time to insight in the future.

6. Summary

The CAP believes that:

- National and International facilities undoubtedly have an important role in the future of HPC provision. STFC should engage fully with development of such facilities in order to shape those facilities both technically and in terms of access models to be optimum for use for STFC science. In so doing we recommend that suitable experts from the communities are engaged in the specification process.
- As in the past, there is still a strong case for HPC support through a coordinated hierarchical provision which makes use of both national and science consortia based machines, each of which provide different capabilities in terms of architecture, cost-effectiveness, and agility.
- It is important to maintain the capability to exploit potential sea changes in the future such as GPUs or developments in Cloud computing.
- An updated strategy should be adopted following consultation with the community.
- CAP concurs with points made in the 2006 strategy review (see Appendix 2), many of which have been implemented. One of those called for *"A further major review of HPC provision should be undertaken by a specially constituted panel in five years time"* . This time has arrived and we recommend that consideration is given to enacting this on the scale of a year.

Appendix 1: Members of CAP

Peter Clarke, Edinburgh [Chair] (Particle Physics Experiment)
Richard Blake, STFC (Computational Science and Engineering)
Jeremy Frey, Southampton (Physical Chemistry & Facilites user)
Neil Geddes, STFC (eScience)
Simon Hands, Swansea (Particle Physics Theory)
Adrian Jenkins, Durham (Astrophysics & Cosmology)
Roger Jones, Lancaster (Particle Physics Experiment)
Ralf Kaiser, Glasgow (Nuclear Physics)
Neal Skipper, UCL (Condensed Matter, Nanotechnology& Facilites user)
Mike Watson, Leicester (Astronomy)

Appendix 2: Executive Summary from 2006 Strategy Report.

1. Computer simulation has become an essential part of PPARC science and is growing in importance in the modelling of complex systems and to confront theory with experiment or observation.

2. High Performance Computing (HPC) is a crucial investigative tool for the theory community across all areas of astronomy (including astrophysics, cosmology and solar system science) and large parts of particle physics. Most of the research activity in the UK theory community relies on access to HPC facilities.

3. The International Review of HPC found that PPARC's HPC programme is producing scientific results that compare with the highest international standards and that some consortia are playing a leading role in setting international standards.

4. The increasing dependency of PPARC science on HPC, which is itself rapidly growing in sophistication and capability, requires that our HPC programme continues to be supported and encouraged to grow. Without it, investment in instrumentation, international subscriptions and manpower will not be able to bear fruit, and our science and technology will lag behind the best (as noted in the 2003 Webber Report).

5. The need for a strategic review of HPC requirements and support was driven by a number of issues: changes in the HPC funding regime; the cross-Research Council International Review of HPC-supported research; the procurement of the new national HPC service, HECToR; rapid technological change and the requirement for increased investment in software development; and the lack of consistent peer-review of HPC projects across PPARC.

6. The PPARC theory community has largely-self organised into consortia to procure, operate and exploit HPC systems. The International Review highly praised this approach which has brought many benefits in the efficient and cost-effective use of HPC systems, in software

development and the training of young researchers. However, there are opportunities for sharing best practice and for greater efficiencies as the HPC landscape evolves.

7. Over recent years, PPARC has benefited from HPC facilities procured with non-PPARC funds through schemes such as SRIF, and (often high) operating costs being absorbed by universities. The introduction of Full Economic Costing (fEC) will significantly increase the PPARC share of the costs of providing HPC facilities either via direct PPARC funding or via charges for access to non-PPARC funded facilities.

8. The consortia require access to a wide range of HPC systems, mostly commodity clusters, but, significantly, also shared-memory systems and a topical machine for lattice QCD. Some of the consortia additionally need occasional access to national class facilities.

9. The funding of software development, in particular the development and implementation of novel algorithms, has been largely neglected to date. This threatens scientific productivity and narrows the hardware options available to consortia, raising costs.

10. PPARC's approach to HPC provision should be holistic rather than piecemeal. The distinctiveness of HPC and commonality of requirements across PPARC applications calls for a programmatic approach in which all aspects of HPC provision are reviewed together.

11. The cost of providing the future HPC facilities and services required to address the immediate science goals of the PPARC research community is estimated at £6.5M p.a. under fEC and the expected value of known bids to the proposed 1st call for proposals in 2006 is estimated to be £15.3M. These figures only include bids expected from the consortia that have made themselves known to the Working Group.

Key Recommendations:

- PPARC should take a programmatic approach to HPC facilities provision and issue three-yearly calls for proposals for HPC facilities and resources starting in 2006. The proposals received should be peer reviewed by PPRP augmented with suitable experts.
- System and software support should be treated as an integral part of funding HPC. The cost-effectiveness, competitiveness and sustainability of the total package of hardware, system and applications software, particularly in relation to international best practice in the field, should be a key part of the peer review.
- An HPC Oversight Committee should be established to have oversight of all HPC projects funded, to foster links with the UK HEC Programme and the National Grid Service (NGS), to disseminate best practice, and to inform future calls for proposals and the peer review process.
- A technical coordinator should be appointed to assist consortia in the preparation of proposals and to provide technical advice to the consortia, PPRP and HPC Oversight Committee.
- PPARC should join the High Performance Computing Studentship scheme initiated by the EPSRC to broaden the training of computational scientists.

- Consortia should be allocated coordination funding to exploit synergies and share experience within and between consortia.
- PPARC partnership in the new national HPC service, HECToR, should be delayed until after the service has been procured and/or the outcome of the HPC peer review is known.
- PPARC should continue to explore with EPSRC options for a joint High-End Computing (HEC) Programme, bringing together some PPARC HPC facilities and HECToR under a common service framework, in order to benefit from shared experience and the opportunity to trade resources, and to position the PPARC community to participate in future national and international HPC developments.
- PPARC should encourage participation by its consortia in the FP7 proposal (HPCEUR) to establish pan-European HPC facilities to ensure that these are best suited to its science needs and that its consortia are strongly positioned to exploit maximally any facilities procured through this mechanism.
- Following each three-yearly cycle of peer review, the PPARC should update the HPC Requirements document and the financial Roadmap.
- A further major review of HPC provision should be undertaken by a specially constituted panel in five years time.

Appendix 3: HPC Sector Details

Summary of HPC computing context

HPC dependent science areas have been subject of many reviews in the past. The most recent strategy was produced in 2006 under PPARC. The executive summary and key recommendations are repeated in Appendix 1. Many of the points mentioned remain valid and are paraphrased below:

1. *PPARC's HPC science is world leading as judged by international reviewers.*
2. *Much, but not all, of the requirement is met by large "Commodity Clusters" with fast commodity backplane interconnects.*
3. *The requirements of the community call for a range of machines with differing attributes.*
4. *The consortia model were praised for cost-effectiveness and training.*
5. *Staff for software engineering is as important as hardware.*
6. *There should be a review of the strategy every three years*
7. *The approach to HPC provision should be holistic.*
8. *There were seen to be opportunities for sharing best practice and further efficiencies as the HPC landscape evolves.*

The many areas of science dependent upon HPC provision are by no means equivalent in terms of the types of codes, its scalability, the number of nodes required, the requirement for inter-processor communication, memory, and the machine architecture.

The majority of the science areas are well served by Commodity Clusters with thousands of cores, with (today) Myrinet or Infiniband backplanes which provide adequate communications speeds. Some applications need nearest neighbour whereas some require longer range communications. Such clusters tend to be very cost effective as there is significant competition in the market and many leading institutions procure such facilities. This model includes several of the Consortia, as well as the STFC facilities.

In the case of STFC facilities there is a "neo-real time" requirement in that the machines need to be guaranteed to be available at the time of experiment runs.

In the case of UKQCD, simulations have been carried out on a dedicated specialised machine (QCDOC) and this has led to the design of the IBM BlueGene series. These machines have historically been very cost effective. UKQCD also utilise Commodity Clusters substantially.

Some consortia are still using large shared memory machines and in some cases only due to legacy code (i.e. if porting effort were available then other architectures would be possible).

A few of the consortia have been able to make use of the national supercomputer (HECToR) and other "traditional" high end HPC This has only been the case when the relatively higher cost-at-point-of-use barrier was removed through awards or other initiatives. This issue remains a barrier to uptake.

The access pattern required for HPC facilities varies during the lifecycle of a science study. In the problem development stages, and for training students and researchers, small scale fast turn-around "agile" facilities are most appropriate. When a simulation is mature, large scale production runs may be carried out on larger scale machines. This is common in the HTC sector as well.

It remains the case today that a "one size fits all" facility is not either architecturally suited to, or cost effective for, all STFC HPC applications, and nor does it provide for the necessary agility.

At the same time it must be recognised that the diverse inconsistencies of the calculation of total cost of ownership and operation of facilities located in central –vs- institutional facilities is a barrier to making a true comparison. The tendency for power, environment and operations staff costs to not be fully accounted into local facility costs at the "point-of-use" makes a very real difference to research budgets upon which the "point-of-use" cost burden would fall. Thus like-for-like overall operating costs are still very difficult to compare.

Finally we note that the traditional boundary between HPC (i.e. specialised machined with often proprietary high speed backplane interconnects optimised for parallel processing) and HTC (i.e. very large commodity clusters with commodity high speed interconnects) is becoming ever more blurred, with traditional HPC applications now able to run on HTC facilities.

Particle Physics Theory : Lattice Quantum Chromodynamics

Description of Computing Task

- The basic physics objective is lattice QCD simulations with physical quark masses in the continuum limit, with small statistical and systematic uncertainty. Other areas studied are QCD under extreme conditions of temperature and/or baryon density, and exploration of QCD-like theories as a dynamical origin of electroweak symmetry breaking.
- Generation of Markov chain ensembles of gluon field configurations
- Measurement of observables on these ensembles: calculation of quark propagators and bilinears, and correlation functions involving quarks and/or gluons.
- Software engineering: code development, maintenance
- Algorithm development
  Training physicists to use code base
- Storage and distribution of gluon configurations and quark propagators, with metadata (QCDgrid)
- Linking to other grids for international data sharing (ILDG)

Underlying Compute Architecture

Gluon Field Generation including the effects of sea quarks requires good memory bandwidth on each node with fast inter-processor communication Depending on the complexity of the process being studied simulations can be either four or five dimensional. The latter allow near exact chiral symmetry required for light quarks and give best control for weak matrix elements such as neutral kaon oscillation and hadronic kaon decays. However five dimensional simulations are more expensive and for other quantities four dimensional approaches are more cost effective.

Communication is nearest-neighbour with occasional global sums. With enough computational nodes, the memory footprint on each node is small and a large cache can be highly beneficial.

The characteristics of QCD simulations are that for a balanced machine each Gflop/s of peak performance should correspond to:

- 1GB/s cache bandwidth per Gflop/s
- 300MB/s memory bandwidth per Gflop/s
- 100MB/s network bandwidth per Gflop/s

In addition few microsecond scale latency is required to obtain many sparse matrix multiplies per second.

Gauge field generation was performed until recently using the specialised QCDOC machines which have a very fast 6-D torus interconnect, and is now migrating to machines such as the IBM BlueGene series having a very similar architecture. There is also substantial use within the community of use tightly-coupled commodity clusters with fast Infiniband interconnect.

Measurement requires the construction of observables using pre-calculated gluon field configurations and quark propagators. It also requires significant computational performance and interprocessor communication for propagator generation, and a large I/O requirement for propagator storage. In many cases it is possible to perform calculations on clusters, benefitting from large memory per node. For problems with smaller memory requirements there is increasing use of GPUs, which offer good performance on sparse matrix-vector problems such as QCD.

UK State of the Art

• Computing
o 10 TFlop/s in UK on UKQCD's 12,000-processor QCDOC in Edinburgh (until 2009); production coordinated with QCDOC machine in Columbia, USA
O(10) Tflop/s on Tightly Coupled Cluster installed in Cambridge as part of DiRAC
O(30) Tflop/s on 2048 node BlueGene/P machine in Swansea as part of DiRAC 13.6 GB/s memory bandwidth per node 5.1 GB/s network bandwidth per node Power: 80KW.
800 TFlop/s on 4096 node BlueGene/Q prototype to be installed in Edinburgh as part of DiRAC 400GB/s cache bandwidth per node 40GB/s memory bandwidth per node 32GB/s network bandwidth per node Power: 400KW
o Access to HPC resources in Europe through eg. DEISA, PRACE, and the collaborations listed below, and USA via HPQCD and RBC-UKQCD collaborations.
o Measurement of observables on (University) clusters with 10^2-10^3 cores, eg. at Edinburgh, Liverpool, Plymouth, Southampton, Swansea

• Memory
o Low per computational core for ensemble generation: eg. on QCDOC key codes use the 4MB on-chip memory of each core
o Physics analysis on clusters benefits from capable nodes with several GB per core

• Storage
o 160 TB parallel (GPFS) file system on BlueGene/P for "scratch" space
o 240 TB distributed RAID system on QCDgrid
We expect the requirement to rise to several petabytes for future machines. Consolidated storage and centralised archival backup facilities is desirable.

• Network
o Reliability is important; our needs do not drive network performance

Roadmap

Year	2011	2012	2013	2014	2015
CPU (Tflop/s peak)	800	800	2000	2000	10000
Disk (TB)	400	600	1000	2000	2000
Tape (TB)	-	-	-	-	-

The sustained performance and memory figures in the table are driven primarily by expected requirements of ensemble generation and measurement of observables via propagator calculation. The balance is anticipated to switch from being ensemble generation-dominated to measurement-dominated as quark masses are reduced and the range of physics studied increases. Multiple systems capable of several hundred Tflop/s sustained performance would suffice to cover the projected UKQCD programme, giving sustained multiple Pflop/s on aggregate.

We will continue to require access to a heterogeneous HPC facility, consisting both of capability resources such as BlueGene, and local clusters, probably University-based, for exploratory smaller-scale projects and the later stages of analysis.

Other points

UK physicists studying lattice QCD organised themselves into the UKQCD Collaboration from the start of the 1990s, to procure and operate the highest performance computing resources, to develop and optimise methods and codes to exploit these facilities, and increasingly to train younger members in their effective use.

UKQCD now comprises scientists from the Universities of Cambridge, Edinburgh, Glasgow, Liverpool, Oxford, Plymouth, Southampton and Swansea. Increasingly, elements of UKQCD collaborate with other international groups to maximise productivity on related but intellectually distinct projects. These links include the UK-USA-Japan RBC-UKQCD collaboration, the UK-USA-Spain HPQCD collaboration, the Swansea-Bielefeld collaboration, and the European Twisted Mass and QCDSF collaborations involving scientists from Italy, Spain and Germany. Shared production of data via international collaboration has become standard practice in LQCD.

Over that last decade or so, UKQCD has developed a significant level of industrial engagement and influence in the supercomputing industry. The joint development of QCDOC with Columbia and IBM led to IBM subsequently pioneering the torus-based BlueGene line, and to an early recognition of the importance of low power computing.
UKQCD is now collaborating closely with IBM Research USA in the development of the BlueGene/Q system, and is currently testing codes on a 512-node protoype.

Use of UKQCD's major resources is determined by a Management Board, informed by collaboration meetings, and reporting to a Project Management Board for the entire DiRAC project: this typically applies to gauge field generation and major analysis (physics measurement) projects. Smaller projects and exploratory simulations are done using small partitions of the major machines, University-based clusters or even local research group resources.

Outlook

There is a trend for modern chips to contain many independent processor cores which potentially offer a huge enhancement of compute-power at relatively little cost. A current example whose use is becoming widespread is the GPU, although limited inter-processor communication speed currently precludes their use in gluon field generation. There will be a need to develop new codes to optimally exploit the full capability of such parallel, multichip systems with multi-core chips. Future machines at the peta- and exaflop scales are likely to consist of thousands of nodes with perhaps hundreds of thousands of processor cores and advanced memory technology, and simulation codes will need to address the need for robustness in the case of failure of one or more. We therefore anticipate a continuing need for expert software support.

International collaboration is becoming more and more important. It is often the case that the UK needs to contribute its share of computing resources to such collaborations. The Grid will likely become more important for sharing ensembles and propagators both within and between collaborations, nationally and internationally, by means of the International Lattice Data Grid (ILDG). Gauge field ensembles and quark propagators are now freely available on the Grid.

Cosmology and Astroparticle Physics: The VIRGO, HORIZON, COSMOS Consortia and the ICG Portsmouth

Description of Computing task

- Fundamental cosmology and the early Universe
- Science exploitation of the microwave sky
- Dark energy and matter content of the Universe
- Modelling large-scale structure and creating mock galaxy catalogues
- Simulations of galaxy clusters and AGN feedback
- Simulating galaxy formation and the intergalactic medium
- Storage of simulation of data and mock catalogues for the "Virtual Observatory"
- Software development (algorithm development for both adding new physics and greater efficiency)

Underlying Compute Architecture

Gravitational simulations, as well as simulations involving radiative transfer, are long range and require significant interprocessor communication. Fast (better than Gbit/s) interconnect or shared memory architectures are necessary for efficient code performance today.

The VIRGO and HORIZON consortia use a relatively small number of MPI codes which run efficiently on large commodity clusters with high speed interconnects. The COSMOS consortium by contrast has a more diverse code base and favours access to very large shared-memory architectures for ease of code development and agilty. The ICG at Portsmouth have requirements that overlap with all three consortia.

Dimensions mid-2011

- Computing
    - Cosmology machine at Durham was upgraded in Nov/10 with 220 iDataPlex nodes each with dual X5650 chips giving a total of 2640 cores and 13Tb of RAM, and 620 Tb of new storage and linked with QDR infiniband interconnect. The upgrade is an addition to an existing machine that has 792 Opteron cores , 2Gb per core, and 300 Tb storage. UK members of Virgo has access through collaborations with non-UK Virgo colleagues to other larger facilities in Europe on a per project basis (e.g. HPCx via DEISA extreme computing initiative, STELLA the LOFAR correlator , hlrb2 at the LRZ Garching, Germany, and Juropa at Juelich, Germany).
    - The HORIZON machine arrived in Oxford in 2010. It is an SGI Altix ICE 8400 EX, consisting of 70 blades each with dual X5650 chips, giving a total of 840 cores and 3Tb of memory and linked by dual rail infiniband, and 120 Gb of storage.
    - The COSMOS supercomputer at Cambridge was upgraded in Aug/10 with an SGI Altix UV1000, consisting of 64 blades, 768 Nehalem-EX cores, 2 Tb of RAM and a NUMAlink5 interconnect, and 64Tb of storage.
    - SCIAMA supercomputer at the Institute for Cosmology and Gravitation at Portsmouth arrived in Jan/11 and has 1048 cores. The compute nodes have dual

X5650 chips and 24 Gbytes of RAM. These are linked by gigabit and QDR infiniband interconnect, and there is 96 Tb of raw storage. SCIAMA was partly funded by SEPNet.

- Memory Typically 2-5 Gbytes per core is sufficient
- Storage
  - Disk - over 1.1Pb of disk in the UK
- Network
  - National network not a limitation. The main issue is that data rates achieved in practice are limited by local networks.

Roadmap

Year	2008	2009	2010	2011	2012
CPU (HPC) [TFlop/s-sust]	15	15	15	30	30
Disk [TB]	600	700	1000	1500	1500
Tape [TB]	100	100	100	100	100

Table showing total across cosmology and astro-particle physics

Other Points

Historically the VIRGO consortium (which is mainly UK led by Durham, but has a significant international involvement) was formed by cosmologists from an n-body simulation background, while the particle cosmologists formed the COSMOS consortium, which is entirely a UK effort led from Cambridge.

Currently both consortia run their own supercomputers. VIRGO by virtue of being an international collaboration has secured time on a number of the largest supercomputers in Europe including HPCx. The Millennium simulation for example was run at the RZG in Garching, Germany.

Both groups, in common with all HPC groups, require agile access to a relatively local system for development and training.

The national facilities, such as HPCx and HECToR while suitable for VIRGO's science programme have proved to be very expensive to access compared to running a consortium based supercomputer. VIRGO has made use of HPCx, but through time awarded in a competition (DEISA extreme computing initiative). The cost effectiveness of the national facilities has also been a major issue for the COSMOS consortium. In addition the COSMOS machine offers a larger globally shared memory than the national systems.

UKMHD Consortium

Description of Computing task:

- Solar and Stellar Dynamos

- Helioseismology, astroseismology
- Solar Corona, sunspots, active regions and solar wind
- Planetary atmospheres
- Star and planet formation

- Galactic dynamos
- Accretion discs
- Relativistic MHD (pulsars, black holes, relativistic jets)

Underlying Compute Architecture
The codes used by the members of UKMHD are all written in MPI and scale very well on clusters with Myrinet or Infinipath interconnect i.e. there is no demand for shared memory or faster communications (HPC).

Consortium: UKMHD (Aberystwith, Bradford, Cambridge, Exeter, Glasgow, Leeds, Manchester, Newcastle, Salford, Sheffield, St Andrews, Warwick)

Dimensions mid-2011

- Leeds: 1920 Westmeres, QDR infiniband, 3.75Tb RAM, integrated in a larger cluster with 3800 cores.
- St Andrews: 1920 Westmeres, QDR infiniband, 3.75Tb RAM.
- Warwick: 1920 Westmeres, QDR infiniband, 3.75Tb RAM.

Roadmap

Year	2008	2009	2010	2011	2012
CPU (HTC) [kSI2k]	0.5	0.5	0.5	0.5	0.5
Cluster (HPC) [TFlop/s-yr]	19	19	24	35	35
Disk [TB]	300	300	400	400	500
Tape [TB]	0	0	0	0	0

UKMHD roadmap.

The capacities in this roadmap are driven by the need to scale up from typically 64 core parallel simulations to 256 core simulations, whilst retaining the capacity to perform smaller runs.

Other Points
The UKMHD Consortium has been in existence since 1996 and its work covers fundamental plasma processes associated with the generation, transport and dissipation of magnetic fields and their associated magnetic energy. It is the correct numerical simulation of the magnetic field that provides the framework for the Consortium.

MIRACLE Consortium

Description of Computing task:

- Modelling of terrestrial and planetary thermosphere-ionospheres.
- Plasma Numerical Simulations.
- Monte Carlo particle transport model for auroral precipitation.
- Calculating the opacity of individual molecules for models of cool stars.
- Modelling of the chemistry and thermal state of species in extra-/galactic clouds/cores.
- Simulating the Radiative Transfer of dust and gas emission from astrophysical objects.
- Simulating the gas and dark matter in clusters of galaxies.
- Producing mock galaxy catalogues for simulating future surveys
- Hydro-simulations of outflows from galactic star bursts.
- Hydro-dynamical simulations of radio galaxy/cluster interactions.
- The structure and stability of dark matter halos.
- The feedback of active galactic nuclei to the dark matter halo.
- The chemical enrichment of galaxies.
- The morphological classification of galaxies.

Underlying Compute Architecture
The consortium today uses large shared memory nodes with fast interconnect between the nodes. In the future this will move to use of a large commodity cluster with Infiniband fast interconnect.

Consortium: MIRACLE (UCL, KCL, Hertfordshire, ICSTM, Manchester)

Dimensions today

- 224 core SUN Cluster (V880,V890) @ 16-64 GBytes RAM, fast interconnect.

Roadmap

Year	2008	2009	2010	2011	2012
CPU (HTC) [kSI2k]	-	-	-	-	-
Cluster (HPC) [TFlop/s-yr]	6	6	6	12	12
Disk [TB]	100	100	100	200	200
Tape [TB]	-	-	-	-	-

MIRACLE roadmap.

Other HPC Areas

There are several other HPC user groups within the UK not explicitly covered in the above. Including Exeter, Leicester, Portsmouth and UcLan. In general the total requirements add approximately 10% of those given in the HPC rows in the tables above.

STFC Facilities: ISIS

Providing adequate computational facilities for data collection and analysis is a major frontier for facilities such as ISIS - all of the science depends on this. It is currently a significant limitation to exploitation. The requirements for each instrument are typically in the medium- rather than high-performance range. So for example resources need to be provided to allow modelling of data as it comes off the instruments.

Experimental time, and hence processing need, is scheduled a long time in advance. This means that use of a shared machine, which is oriented to the neo-real time Facility needs, would be most efficient.

The number of users at ISIS is very large - typically two groups per instrument per week. Instrument scientists can provide support for the use of instrument and data collection software and of course interpretation, but in many cases it is not possible to train the instrument scientists in state the of the art modelling and interpretation software. Notable successes have come when the Facility has specialised staff who can take responsibility both for this software development and its use - for example the C-Lab in Grenoble, or the CSED staff developing ab-initio chemistry codes who are physically based within ISIS. ISIS could benefit from greter synergy between its computational and experimental
capabilities.

Data storage and transfer needs for ISIS and CLF are not large by modern standards. But some Diamond data is produced too fast to be transferred elsewhere (e.g. Daresbury) for initial processing - this needs to be done locally at RAL.

STFC Facilities: Central Laser Facility Plasma Physics Modelling

The CLF Plasma Physics group has three core computing objectives to fulfil:
1) Support the CLF user base (particularly to aid publication of experimental results),
2) Support the HiPER project, and
3) carry-out our own cutting-edge research.
The use of parallel computing resources is critical to all three.

Description of Computing task

1. Development of an extended hydrodynamic code using Adaptive Mesh Refinement grid
2. Vlasov Fokker Planck codes for energy transport physics studies
3. Hybrid-PIC code development using quasi-static approximation for laser and beam driven particle accelerators
4. Hybrid-PIC code development for fast electron energy deposition in dense plasmas
5. Simulation programme to support CLF users
6. Development of Eulerian Vlasov solvers
7. Simulation programme to support HiPER project

8. Training of PhD students (CLF Users) in use of code base
9. Visualisation support (e.g. successor to VIS Cluster)

Underlying Compute Architecture

Parallel simulations require computer architecture comparable to SCARF LEXICON II or better. The machine already has more than enough users to fully occupy the available resources. So far, it has served STFC well in two-dimensional simulations, but is struggling to accommodate three-dimensional problems which are at the cutting edge of research. Doubling the size of this cluster would allow three-dimensional problems to run alongside existing simulations programmes.

Dimensions Today

- SCARF-LEXICON II – 544 cores
- Memory – 3GB per core
- Disk requirements: 2-3TB per user (8 users in total)
- Infiniband network
- SCARF LEXICON I is used for code development, testing problems and external collaborators.

Year	2011	2012	2013	2014	2015
Number of Cores	544	1088	1088	1088	2176
Disk [TB]	30	60	60	60	120
Tape [TB]	-	-	-	-	-

Challenges
Our aims in code development, support, and research are extensive and ambitious. Clearly carrying all of these out in the next 2-3 years with current manpower is difficult. Progress could be accelerated by provision of additional personnel effort.

Other Points
Training and parts of code development are carried out in partnership with universities, under the umbrella of the CCPP and EPSRC funded research grants.