Survey of Users' Needs

A joint scientific and technological study undertaken by
CINECA and CSC for the ENACTS network

User Requirements Report

Giovanni Erbacci and Claudio Gheller (CINECA)
Satu Torikka (CSC)

December 31, 2004

http://www.enacts.org

Table of Contents

1 Overview

1.1 The ENACTS Project

ENACTS is a Co-operation Network in the 'Improving Human Potential Access to Research Infrastructures' Programme. This Infrastructure Co-operation Network brings together High Performance Computing (HPC) Large Scale Facilities (LSF) funded by the DGXII's IHP programme and key user groups. The aim has been to evaluate future trends in the way that computational science will be performed and the pan-European implications. As part of the Network's remit, it has run a Round Table to monitor and advise the operation of the four IHP LSFs in this area, EPCC (UK), CESCA-CEPBA (Spain), CINECA (Italy), and BCPL-Parallab (Norway).

This co-operation network follows on from the successful Framework IV Concerted Action (DIRECT: ERBFMECT970094) and brings together many of the key players from around Europe who offer a rich diversity of High Performance Computing (HPC) systems and services. In ENACTS, our strategy has involved close co-operation at a pan-European level - to review service provision and distil best-practice, to monitor users' changing requirements for value-added services, and to track technological advances. In HPC the key developments are in the area of Grid computing and are driven by large US programmes. In Europe we urgently need to evaluate the status and likely impacts of these technologies in order to move us towards our goal of European Grid computing, a 'virtual infrastructure' - where each researcher, regardless of nationality or geographical location, has access to the best resources and can conduct collaborative research with top quality scientific and technological support. ENACTS has provided participants with a co-operative structure within which to review the impact of Grid computing technologies, enabling them to formulate a strategy for increasing the quantity and quality of access provided.

Table 1.1 Goals and Membership - ENACTS participants by role and skills

Centre	*Role*	*Skills/Interests*
EPCC	IHP-LSF	Particle physics, Materials science
ICCC Ltd	User	Optimisation techniques, Control engineering
UNI-C	LSF	Statistical computing, Bioinformatics, Multimedia
CSC	User	Meteorology, Physics, Chemistry, Biosciences
ENS-L	Society	Computational condensed matter physics, Chemistry
FORTH	User	Computer science, Computational physics, Chemistry
TCD	User	Particle physics, Pharmaceuticals
CINECA	IHP-LSF	Meteorology, Computational Physics, Astrophysics, VR
CSCISM	User	Molecular sciences
UiB	IHP-LSF	Computational physics
PSNC	User	Computer science, Networking
UPC	IHP-LSF	Meteorology, Computer science
NSC	User	Meteorology, CFD, Engineering
ETH-Zurich	LSF	Computer science, Physics

1.2 *The ENACTS Co-operation Network*

Four of the participants (EPCC, CINECA, CESCA-CEPBA and Parallab) are LSFs that have provided Researchers' Access in HPC under the HCM and TMR programmes and they have co-operated more closely in the Transnational Access programme.

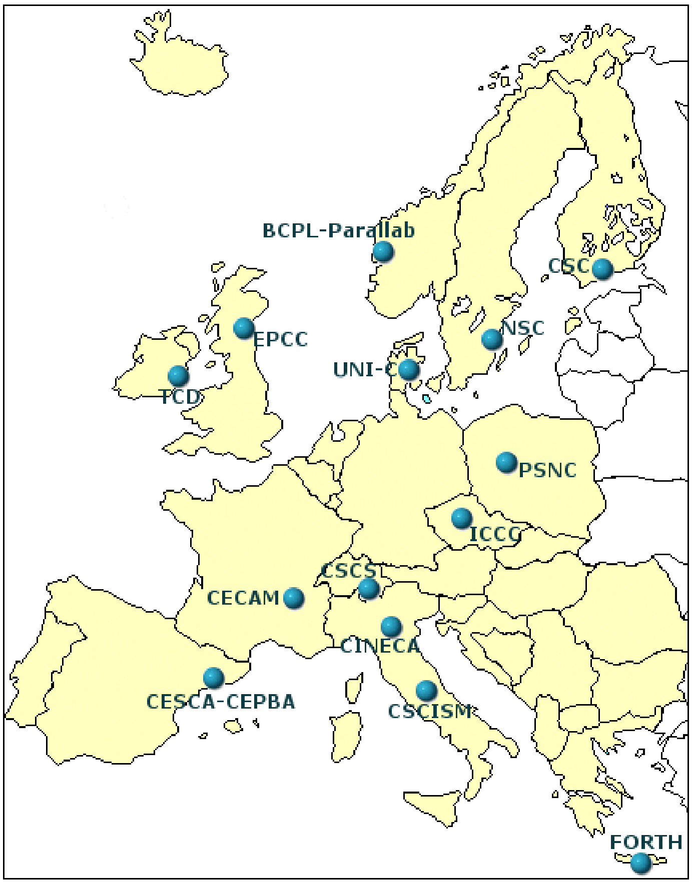

Figure 1.1 ENACTS members

These four LSFs have provided access to over 500 European researchers in a very wide range of disciplines and have been thus well placed to understand the needs of academic and industrial researchers. The other 10 ENACTS members have been drawn from a range of European organisations with the aim of including representation from interested user groups and also by centres in economically less favoured regions. Their input has ensured that the Network's strategy has been guided by users' needs and relevant to smaller start-up centres and to larger more established facilities.

See http://www.enacts.org for up-to-date information on the ENACTS project.

1.3 Joint Scientific/Technological Activities and Studies – Survey of Users' Needs

The present report aims at presenting the results of the "Survey of Users' Needs" scientific/technological activity, whose objective are to determine users' requirements for access to HPC facilities and Datastores and assess the implications for changes in their working patterns if these were provided within a metacentre model. Furthermore, the report outlines how users perceive emerging technology affecting their research and looking at the technological barriers to mobility of researchers. The results are based on the opinions of both large user groups and of individual users of high performance and distributed computing facilities in Europe.

1.3.1 Work Plan

One of the key tasks for ENACTS is to collect information from existing and potential users of High Performance and distributed computing about their future requirements. While much of the qualitative, in-depth information on requirements can be collected via ENACTS participants (both HPC centres and user representatives), ENACTS also aims to collect quantitative information from a wider range of groups via a web based questionnaire. This will enable the Network to check the requirements of a far wider cross-section of the computational science community than would otherwise be possible.

The questionnaire (see the details from Chapter 2) has been designed to gain information in a range of areas including the value placed on current services, the limitations and applications bottlenecks, the level of user experience and expertise and future requirements.

In addition, detailed information has been collected from key users groups, represented or identified by ENACTS participants, by means of an interview. Each user group was asked the same series of open ended questions about their requirements.

1.3.2 Main Tasks

The "Survey of Users' Needs" activity consists of five workpackages, totalling 6.6 staff-months of effort. The elapsed time for the activity is 6 months. The workpackages are summarised below.

Table 1.2 User Survey Workplan

	Workpackage	Effort	Produced By	Accepted By
WP1	User survey	2.0	CINECA, CSC	Network Co-ordinator
WP2	Results analysis	1.0	CINECA, CSC	Network Co-ordinator
WP3	In-depth interviews	2.0	CINECA, CSC	Network Co-ordinator
WP4	User requirements report	0.5	CINECA, CSC	Network Co-ordinator
WP5	Dissemination	0.6	CINECA, CSC	Network Co-ordinator
WP6	Project Management	0.5	CINECA	

The Management Committee of the ENACTS project has performed quality assurance on project deliverables and has accepted them on behalf of ENACTS. The Study Co-ordinator has agreed a detailed project and dissemination plan with the Network Co-ordinator who monitors project progress on behalf of ENACTS. The Study Co-ordinator has communicated with participants via e-mail and made deliverables available for comment and review via internal project web pages.

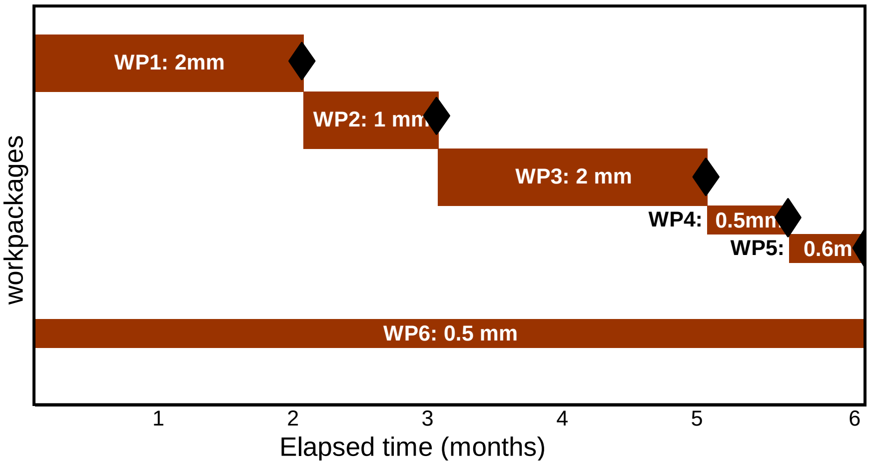

Figure 1.2 Gantt Chart for the User Survey Study Project

WP1: This is a survey of the current experience and requirements of users and potential users of HPC facilities. It involves the design and promotion of a web-based questionnaire for users of European LSFs and HPC centres. This workpackage comprises 2 staff-months of effort and has been performed by CINECA and CSC. The other HPC centres in the ENACTS network have promoted the questionnaire to their users and contact organisations.

WP2: The questionnaire returns have been analysed and data-mined to look for significant trends. This workpackage has taken 1 month of effort and will be undertaken by CINECA and CSC.

WP3: Eleven in-depth phone or face-to-face interviews have been conducted with representatives of significant computational science research groups or organisations in Europe, to solicit their views and opinions on future requirements for HPC. This activity has been led by CINECA and CSC, with assistance also from other ENACTS participants, who have identified target interview groups and conducted interviews. This workpackage took 2 months of effort.

WP4: The output from the analysis of the user survey and the completed in-depth interviews has been used to produce the User Requirements report (this report). This report details the differing requirements of significant user groups in Europe and summarises user requirements for Grid Computing. CINECA and CSC have written this report. The workpackage had a duration of 0.5 staff-months.

WP5: All the participants are required to publicise the availability of the User Requirements report to interested parties. This workpackage has a duration of 0.6 staff months.

WP6: Project Management. CINECA has undertaken this activity. This workpackage required 0.5 staff-month of effort, over the 6 elapsed months of the project.

1.3.3 Involved institutions

This activity has been undertaken by CINECA and CSC.

CINECA, Italy (http://www.cineca.it)

CINECA is an Inter-University Consortium, composed by 24 Italian Universities and CNR (the Italian National Research Council), founded in 1969 to make the most advanced tools for scientific calculation available to academic users and public and private researchers. Since then, the consortium acquired the most powerful and state-of-the-art high-performance computer systems available on the market, investing in equipment for high performance computing which has proven to be particularly effective and of great commercial success.

Its advanced policy in continuously up-dating the available equipment, as well as the high degree of professional skill in management and user support have allowed for a high degree of technological transfer to both public and industrial environments, and a service to a vast community of users, spread out over the entire national territory. The usability of the service is guaranteed by a capillary and a powerful network, continuously maintained at the state-of-the-art.

The Consortium computing centre acquired powerful state-of-the-art supercomputers and steadily invests in high-performance computing systems. Its top line supercomputers currently include an - IBM SP Power4 Regatta with 512 processors, 2.67 Tflop/s peak-performance), and a Pentium IV Linux cluster with 768 processors assembled by IBM. This heterogeneous and balanced set of resources is completed by VIS.I.T - Visual Information Technology Laboratory, featured a Virtual Theatre and equipped with SGI Altix 64 processors graphics supercomputer, and other Virtual Reality (VR) and video-animation devices.

The specific skills include vector and parallel optimisation of algorithms and code development, in co-operation with scientists from a wide variety of scientific fields: physics, computational chemistry, fluid dynamics, structural engineering, mathematics, astrophysics, climatology, medical image applications and more. Moreover, CINECA is involved in several EC funded projects, in the field of research and technology transfer.

CSC, the Finnish IT Center for Science (http://www.csc.fi)

CSC is the Finnish IT center for science, governed by the Ministry of Education. CSC develops and provides modelling, computing, and information services for universities, polytechnics, research institutions and industrial companies. CSC started its operations in 1971 at the Finnish State Computer Center as the support unit for

Univac, the first mainframe computer of Finnish universities. In 1989, CSC acquired the country´s first supercomputer for the needs of the academic research community.

Through CSC, researchers have access to Finland´s widest selection of scientific software and databases, together with leading-edge supercomputing architectures. All this is made available by the high-speed Funet (Finnish University and Research Network) telecommunication connections, maintained by CSC. CSC maintains the largest selection of application programs in Finland. More than 200 different applications have been installed in CSC’s computing environment. CSC offers high-quality expertise in various scientific disciplines such as chemistry, biosciences, earth sciences, physics, statistics, electromagnetics, computational fluid dynamics, structural analysis, and mathematics as well as scientific visualization. A large collection of scientific databases is installed on CSC´s computers and are available to researchers. Furthermore, CSC maintains an index database server for scientific libraries and a database describing the operation of universities.

CSC’s high rating is based on its versatile supercomputing environment with total computer capacity 2.5 Teraflop/s. CSC's most powerful computer IBM eServer Cluster 1600 is capable of computing 2200 billions floating point operations per second. It equals 400 Pentium P4-processors. CSC's newest system, now being installed, consists of two SunFire 15K servers. The final configuration, which will be available in early 2005, has 96 UltraSPARC IV processors and 384 gigabytes of memory. The new system will offer a comprehensive set of scientific applications to CSC's customers, and will eventually replace the current SGI Origin 2000 application server.

CSC’s experts participate in research projects of universities and industrial companies. This ensures the free flow of scientific computing information and know-how between the academic research world and the Finnish industry. Thanks to research collaboration, research projects have been eligible for funding from the European Union, the Academy of Finland and the National Technology Agency.

1.3.4 The Authors

Dr. Giovanni Erbacci

Giovanni Erbacci holds a laurea degree in Computer Science from the University of Pisa. Since 1999, he has co-ordinated the Scientific Computing Group, in CINECA's High Performance Systems Division. He is active in promoting High Performance Computing, co-operating both with academic and industrial institutions. Moreover he is responsible for CINECA's training and education activities in the field of HPC. Since 1992 he has organised and directed CINECA's Summer School on Parallel Computing. He is a contract professor of Parallel Computing at the University of Ferrara, and he has been the advisor of graduation thesis on the topics of parallel computing. His main interests include Parallel Algorithms and Architectures, Parallel Programming, Parallel Languages and Compilers and Performance Evaluation of Parallel Systems and Algorithms. Giovanni Erbacci is the author or the co-author of more than 30 papers published in journals and conference proceedings and he is a member of ACM and the IEEE-Computer Society. He has large experience in EC projects. These include the participation in a range of projects under the HPCN

NOTSOMAD Technology Transfer Node (1997 - 2000). In Framework IV he was involved in the ICARUS Programme and in the DIRECT Concerted Action. In Framework V he was the Project Manager of the TMR-Access MINOS programme (2000-2003) and he participates to the ENACTS Co-operation Network in the IHP Access to Research Infrastructures Programme, where he is on the managing committee. Currently he is the scientific coordinator of HPC-Europa, the EC Support for Research Infrastructure Programme.

Dr. Claudio Gheller has a degree with honour in Physics from the University of Padova, and a Ph.D in astrophysics at SISSA/ISAS in 1997. From December 1997 to May 1999 he had a Post-doctorate position at CINECA working on numerical methods for astrophysics. Since May 1999 he is member of the staff of the Supercomputing division at CINECA and he is in charge of the astronomical research support. He collaborates with various astrophysical in the field of cosmological simulations, data management and visualization, Virtual Observatory. He was one of the responsibles of the EU project ESTEDI and COSMO.LAB (IST – 5FW Programme).

M.Sc. (Tech.) Satu Torikka works as a Research Coordinator at CSC since October 2001. Her work focuses on cooperation projects on High Performance Computing, and support of computing customers. Grid and ENACTS are her current main activities. Her studies are derived from the laboratory of environmental engineering at the Helsinki University of Technology. Before joining CSC she has worked for Vaisala Ltd. concentrating on technical documentation and customer support of measurement systems for meteorology and environmental sciences, and Razorfish Ltd., a provider of digital solutions for the Internet and mobile technologies.

1.4 Report Content

The report consists of 7 sections

- Chapter 1 introduces the ENACTS project and the objectives of the specific activity.

- Chapter 2 describes the structure and the content of the questionnaire. Furthermore it presents the data collection and analysis procedure adopted.

- Chapter 3, 4 and 5 present the details of the various sections of the questionnaire and the results obtained for each section.

- Chapter 6 presents the result of in-depth interviews with several selected users.

- Chapter 7 draws the conclusions and proposes some recommendations and suggestions for the future development of users needs driven research tools.

1.5 Dissemination Activities

The final version of the User Requirements report is now available on the ENACTS web site. An email has been sent to all those researchers and companies who kindly participated in this survey. A one page flyer has been designed by CINECA and CSC and printed by EPCC.

CINECA and CSC have made several presentations about the findings of the Survey of Users' Needs. Satu Torikka presented findings in the questionnaire on 13th October 2004 at the Grid Seminar in Espoo, Finland, that was targeted at CSC staff and the national Material Sciences Grid partners. Giovanni Erbacci and Satu Torikka made a presentation at the workshop on Grid computing on 16th and 17th November 2004 in Lyon, France. The workshop, organised by Cecam, was an encounter between users and operators of HPC centers in Europe.

1.6 Acknowledgements

The authors of this report would like to thank all the people that took part to the survey, filling the questionnaire. Moreover, we would like to express our gratitude to the representatives of the significant computational science research groups who accepted to be interviewed:

Axel Berg (SARA Amsterdam)
Luca Biferale (INFM Rome)
Gianluigi Bodo (Astronomical Observatory of Turin)
Roberto Capuzzo Dolcetta (University of Rome "La Sapienza")
Jean-Christophe Desplat (EPCC, University of Edinburgh)
Jari Järvinen (CSC - Finnish IT center for science)
Kari Laasonen (University of Oulu)
Matthias Mueller (HLRS, University of Stuttgart)
Risto Nieminen (Helsinki University of Technology)
Kai Nordlund (University of Helsinki)
Sven Stafström (Linköpings Universitet)

2 The Questionnaire

2.1 Questionnaire Contents

The contents and design of the questionnaire and the set up of the users survey took place from January 20, 2004 to the beginning of April 2004, when the draft version of the questionnaire was presented to the ENACTS partners. The final version of the questionnaire corrected according to the partners' suggestions and recommendations was completed on May 19, 2004. From this date up to June 16, 2004 the questionnaire was officially open and data collected.

You can find the questionnaire in Appendix 1. The questionnaire consists of 45 questions. Most questions are multiple answer questions. In addition, some space is left to the participants to provide their considerations and suggestions. The questionnaire is divided in four sections:

1. **Introduction –** In this section participants were asked to provide information about their professional background and scientific interests. Furthermore, two questions investigated the typical software usage of the participant and his/her research group. Please see Section 2.4 Introductory Information on Questionnaire Participants, page 15.

2. **Computing Environment Needs –** This section has the goal of investigating the current working environment of the user, with respect to computational resources. The first set of questions is dedicated to the personal computing environment of the user, considering the operating system, the network, the connection to the HPC infrastructure, focusing on the quality of the service, in terms of stability, efficiency, support etc. The typical usage of computational resources and the needs for HPC or distributed computing are investigated. In particular, we try to find out the most important needs associated to computational resources and the bottlenecks in the present computing environment. Finally, the participants assessed how their needs will have changed by the year 2007 with respect to their computing environment.

3. **Application Needs –** The third section is specifically dedicated to the applications that are typically adopted in his/her research activity by the user. The first questions are related to the usage of the computing systems (CPU intensive vs. data intensive), the programming languages (Fortran, C, C++ etc.), the libraries and the compilers required by the application programs. A couple of questions analyse the way data are stored and managed (specific standard formats, database usage). Then, the user is required to specify which are the most important tasks in his/her HPC work (code development/porting, data management/analysis etc.), if his/her application can run on parallel/distributed architectures and, if so, which is the adopted parallel computing paradigm. The last question investigates if the user would be willing to spend some time to port his/her application on a distributed environment.

4. **Grid Environment Experience and Future Needs –** The last section focuses on the Grid technology. In the first half, the questions have the goal of investigating the present interaction of the user with the Grid in its various aspects, like collaborative work, or specific Grid-enabled applications (Virtual Observatory, Access Grid etc.). Furthermore the level of knowledge of the available middleware (Globus, Condor, Unicore etc.) is asked. Finally the user is required to specify the characteristics of Grid environment he/she thinks are critical (security, web-based access, high bandwidth network etc.) and which are the present main difficulties for the interaction between the Grid and end-users. The second part of this section is dedicated to analyze the possible contribution of the user to the development of the Grid, in terms of research, technology, software or data

The survey was completed by in-depth interviews to representatives of significant computational science research groups or organisations in Europe. The interviews are presented in detail in Chapter 6.

2.2 Data Collection and Analysis of the Results

The questionnaire was officially closed on June 16th, 2004, with a total of 125 replies. Html and php were used in the questionnaire to automatically save the collected data, with interface to mySQL. The statistical package SAS was used for analysing the results.

2.3 Questionnaire Participants

A large population from the scientific community across Europe was invited to take part in this study. The people invited to fill the questionnaire were selected such that they represent real (everyday) HPC and/or Grid users, in particular researchers and scientific software developers. A further requirement was to focus on medium-big computational resources users, so that they can provide a meaningful feedback on high-end CPU devices and infrastructures. The ENACTS partners suggested the names of users that could be interested in answering the questionnaire, an average of 50 users per country. In addition, the questionnaire was advertised to specific interest groups on the Grid (Global Grid Forum users, NorduGrid users) and to contacts of the 6th EU Framework Programme Grid projects (DEISA, EGEE).

The following table summarizes the number of answers by each European Country, totalling 125 participants.

The answers come from all over Europe, including the new EU member states. The North Europeans, Sweden, Finland, Norway and Ireland, represent the biggest number of participants. The group Other contains the participants who did not tell the country where they are working.

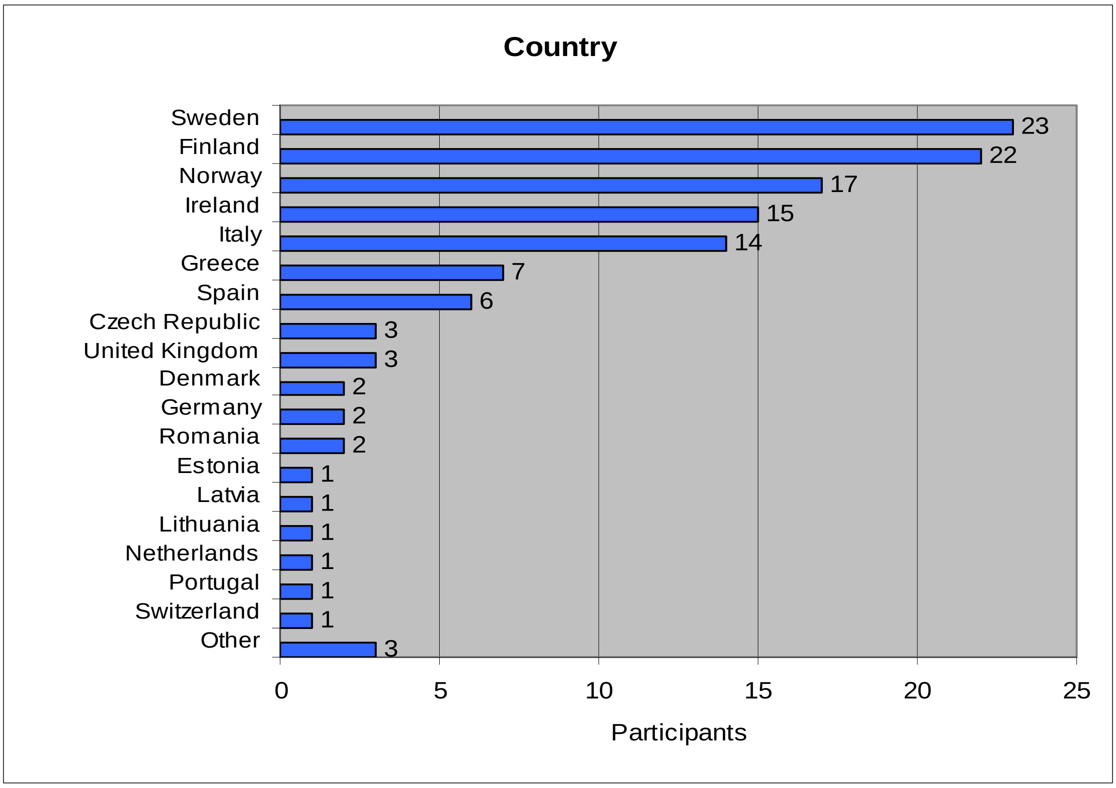

Figure 2.1 Countries where the participants are working (125 participants in total)

2.4 Introductory Information on Questionnaire Participants

The participants were asked about their personal background and work details, scientific background and the usage patterns in their work. Most participants come from universities (71%) and research institutes (24%). Users from industry are also represented in this survey.

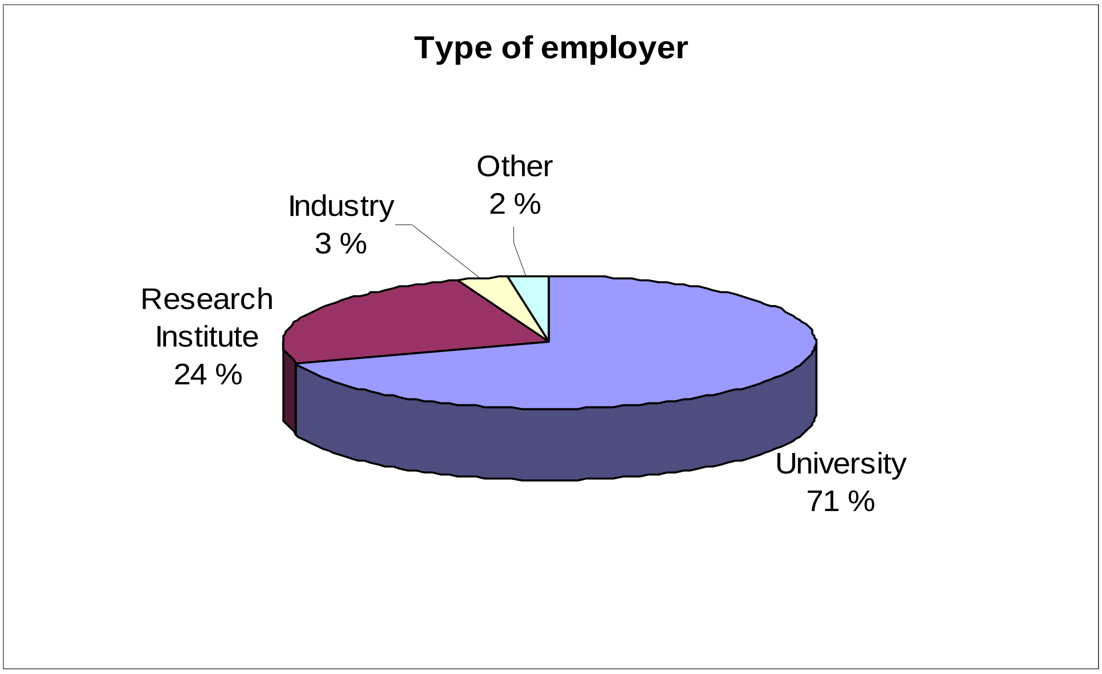

Figure 2.2 Participants typically come from universities and research institutes

Researchers form the largest group of participants (45% of the total). The second biggest group is PhD students (27%), followed by group leaders (17%), code developers (4%) and engineers (1%). Other positions (6%) are professors, associate professors, system administrators and students.

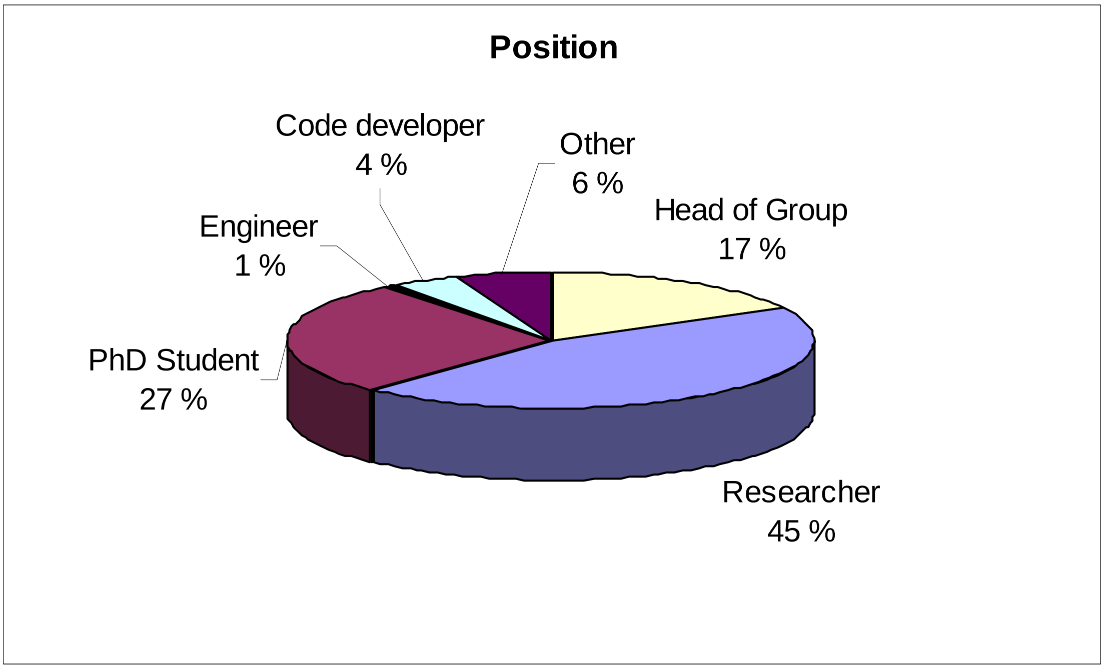

Figure 2.3 Position of participants

Half of the participants' groups are comprised of people from one institution only. However, 20% of participants cooperate at international level in Europe, and 15% have members from outside Europe in their groups. Also national collaboration is common, 16% of participants' groups have members from other institutions.

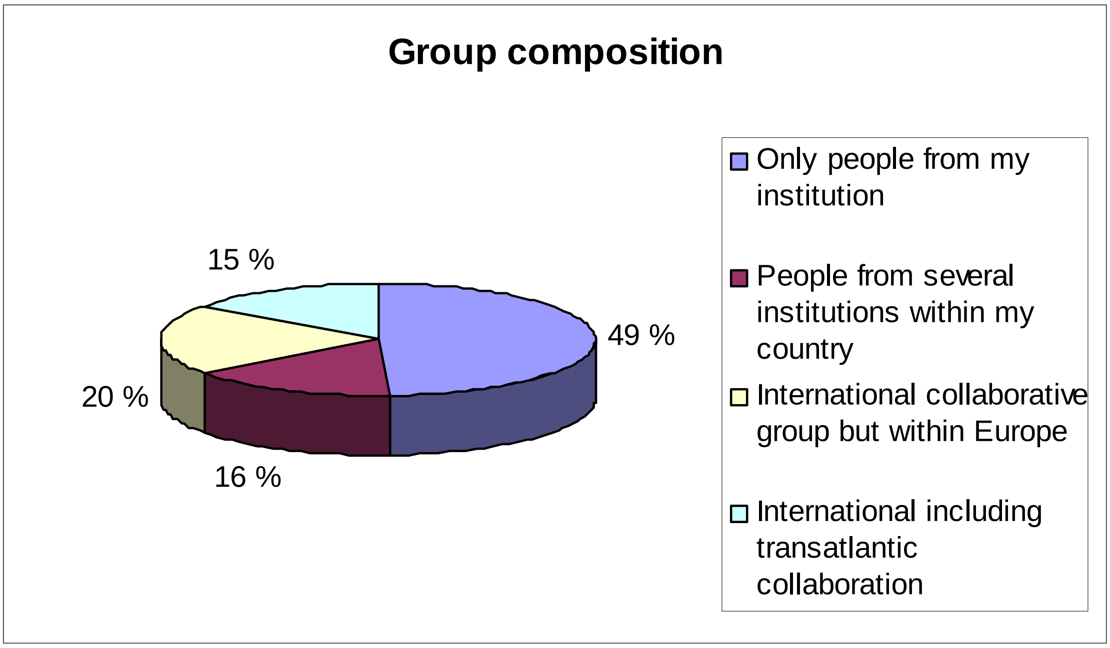

Figure 2.4 Group composition of the participants

The size of the participants' groups comprised of medium sized groups (groups sized between 2 and 10 people, 62%) and larger groups (groups sized between 11 and 50 people, 31%). 3% of groups were very large, comprising over 100 members. 2% of groups had only one member.

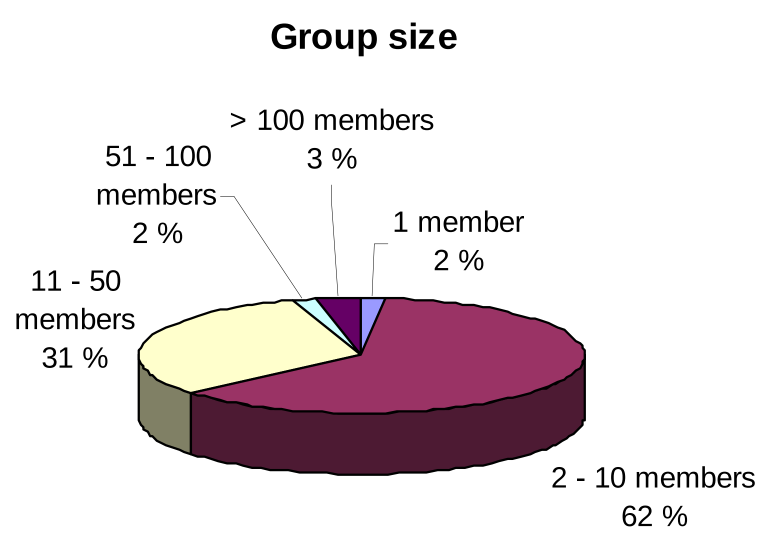

Figure 2.5 Group size of participants

2.4.1 Scientific Profile

As shown in Figure 2.6, a big number of participants are involved in computing sciences; 26 participants out of 125, i.e. 21% of the participants. The chemists come next (16% of the participants), followed by the condensed matter physicists (14% of the participants). The physicists (condensed matter, astro-, geo-, high energy, and other physicists) represent altogether 38% of the participants.

The work of 73% of the participants is multidisciplinary while the work of the rest of the participants (27%) is not. Among the secondary fields, computing sciences and condensed matter physics are the biggest.

In addition, participants were asked to tell about their research interests. 5% of the participants expressed that the Grid is their main research interest; they were mainly computing scientists. Computing scientists also mentioned parallel computing as their research interest. The main research interests of the chemists are quantum chemistry, computational chemistry and molecular dynamics. Environmental scientists have climate, atmosphere, ocean and ice modelling as their research interests. Bioscientists' interest is directed towards bioinformatics.

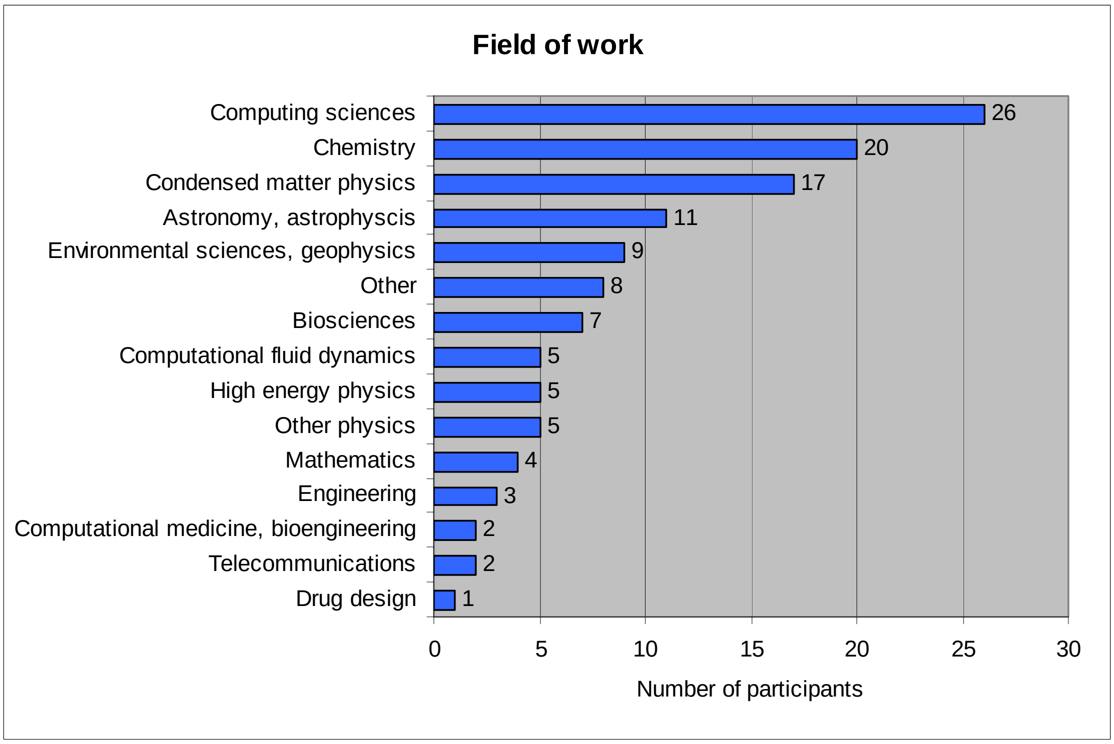

Figure 2.6 Field of work of the participants' groups

2.4.2 Usage patterns of computers

Of the three profile definitions, the participants selected the approximate percentages that best described their usage of computers at work (Figure 2.7). Since the users could have all these profiles, they have been asked to specify which fraction of their job includes:

- Running pre-compiled applications
- Modifying the source code when needed
- Code development and compiling or modifying the original source code.

Figure 2.7 shows that 18% of participants are pure code developers, and 10% use entirely pre-compiled applications. Source code modification forms a part of most participants' work, but it is seldom the main job.

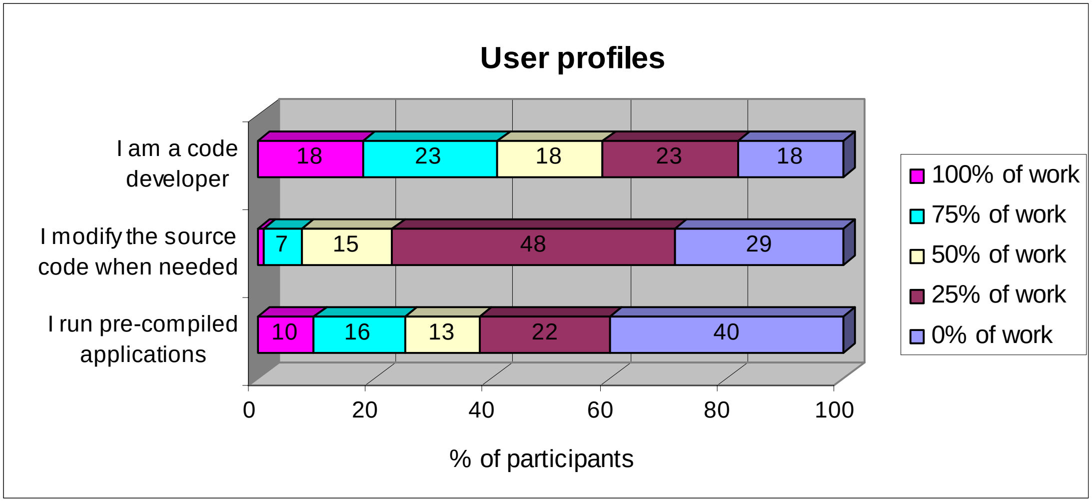

Figure 2.7 User profiles

In addition to personal profile definitions, the participants selected the approximate percentages that best described the usage pattern of their group, totalling 100% (Figure 2.8). Figure 2.8 shows that 10% of participants' groups are pure code developers. None of the groups uses entirely pre-compiled applications.

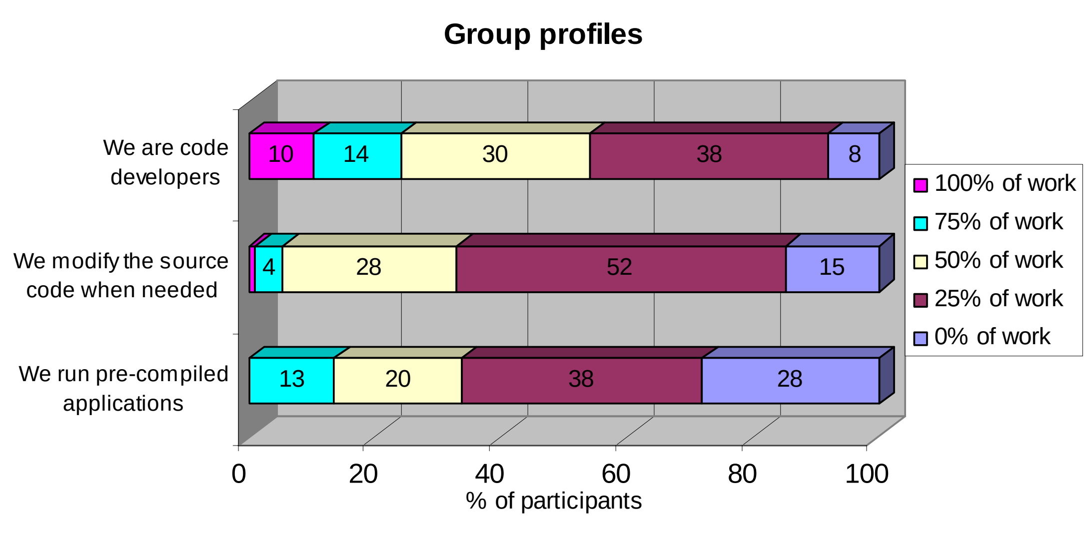

Figure 2.8 Group profiles

When comparing Figure 2.7 to Figure 2.8 you can notice that a typical respondent of this survey is a code developer. Groups typically consist of both code developers and users of pre-compiled applications.

3 Computing Environment Needs (Questionnaire Part 2)

This section has the goal of investigating the current working environment of the user, with respect to computational resources. The first set of questions is dedicated to the personal computing environment of the user, considering the operating system, the network, the connection to the HPC infrastructure, focusing on the quality of the service, in terms of stability, efficiency, support etc. The typical usage of computational resources and the needs for HPC or distributed computing are investigated. In particular, we try to find out the most important needs associated to computational resources and the bottlenecks in the present computing environment. Finally, the participants assessed how their needs will have changed by the year 2007 with respect to their computing environment.

3.1 *Analysis of the Results*

Linux is the most commonly used operating system among the participants (114 users out of 125). Unix comes next (73 users), followed by Windows (59 user). MacOS has 13 users (Figure 3.1). Other operating systems mentioned are Solaris and BSD and freeBSD.

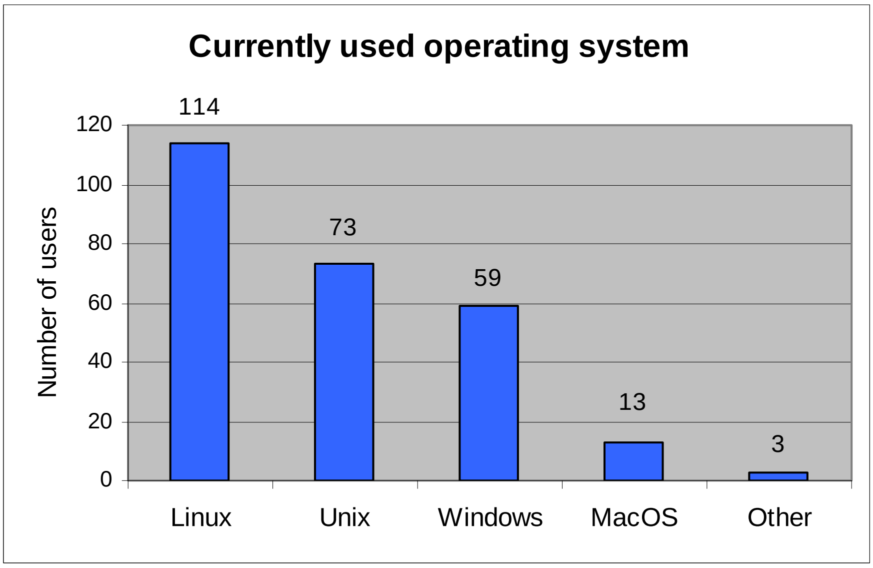

Figure 3.1 Operating systems used by the participants

Nearly every participant of this survey has access to Local Area Network (118 users out of 125). Only 18 users out of 125 use a wireless connection (WLAN). ADSL (XDSL), modem and ISDN are rare (Figure 3.2).

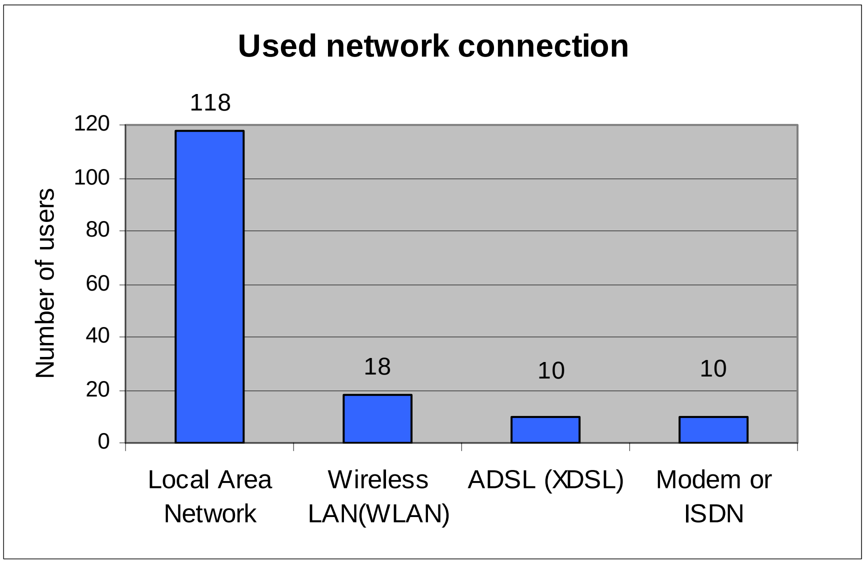

Figure 3.2 Network connections used by the participants

The participants typically connect to their computing environment via ssh (116 users out of 125). Of those, 48 use X windowing. 25 users have a client on their own machine while 15 users connect via www (Figure 3.3).

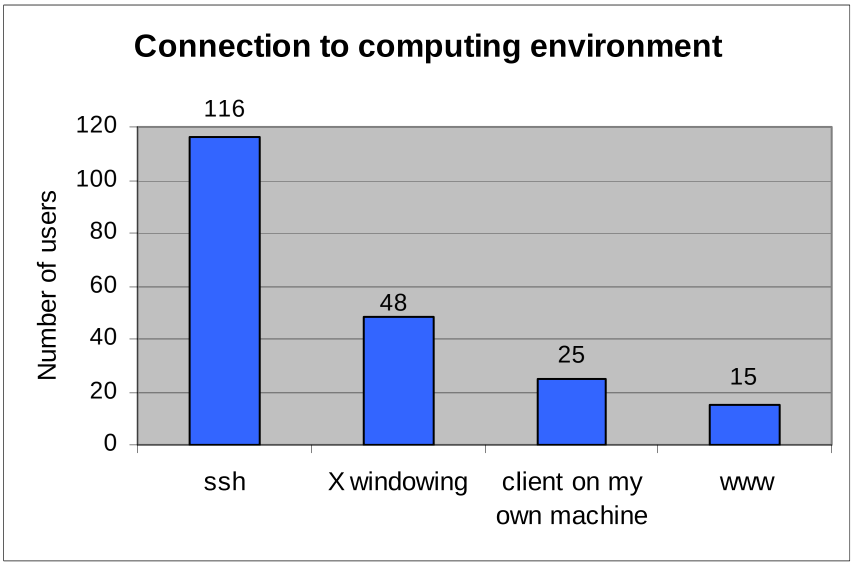

Figure 3.3 Connection types used by the participants

The primary access to the computing environment is local, inside home institution (95 users out of 125). Access outside own institution but within home country is fairly common (51 users). 14 users have access to abroad as their primary access to the computing environment (Figure 3.4)

The participants who selected Access to abroad as their primary access gave detailed information on their computing environment. They use supercomputing facilities located in CERN, Nordic countries (via NorduGrid), USA (e.g. NERSC), United Kingdom, Estonia and Switzerland. Also, CVS repositories are used.

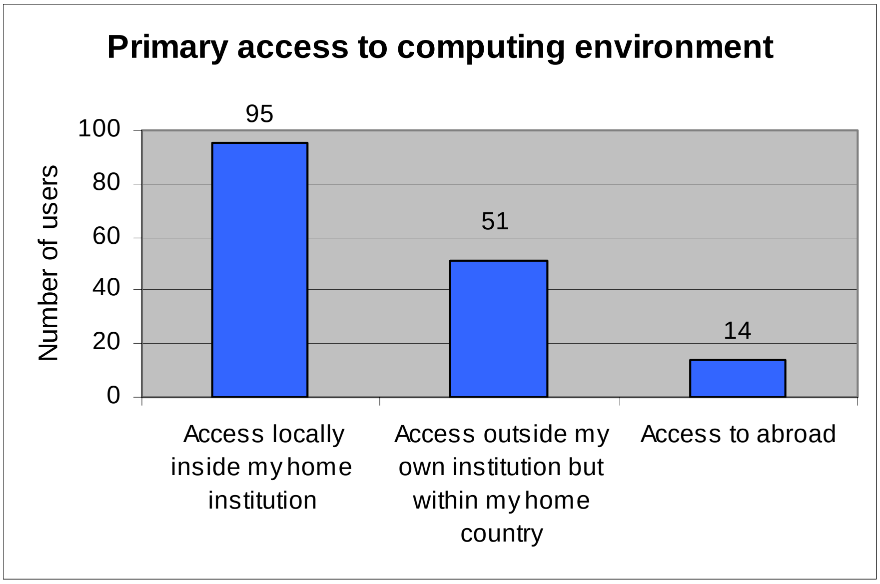

Figure 3.4 Primary access to computing environment

Figure 3.5 shows that 38% of participants use a bandwidth of 100-999 Mbps when connecting from their own machines to their computing environment. 14% of participants have a bandwidth of 1-3 Gbps, and 3% a bandwidth bigger than 3 Gbps. 24% of participants do not know the bandwidth they use.

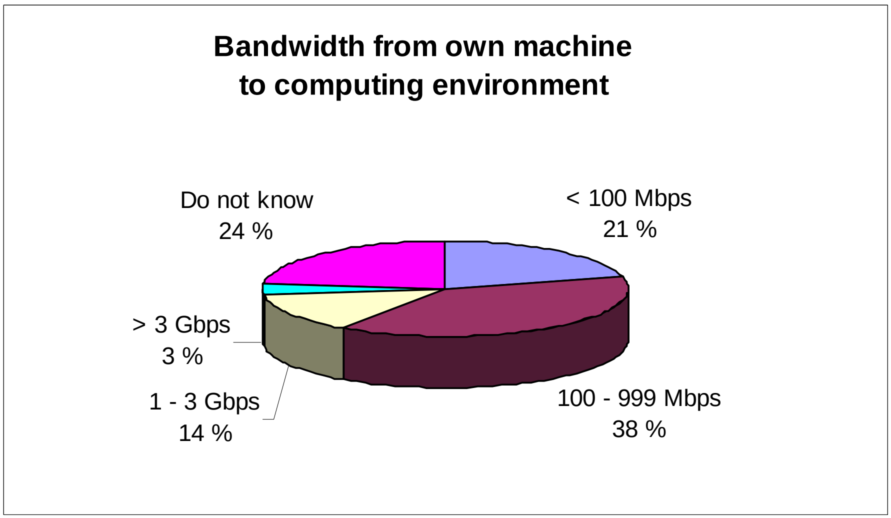

Figure 3.5 Bandwidth to users' computing environment

The computing services were evaluated on the scale of 1-5, where 1 is poor and 5 good. Figure 3.6 shows that most of the participants can have a high quality computing environment available. Computers are easy to use, and the information about computer downtime periods, like service breaks, are sufficient. The quality of the user support is also considered to be at high level.

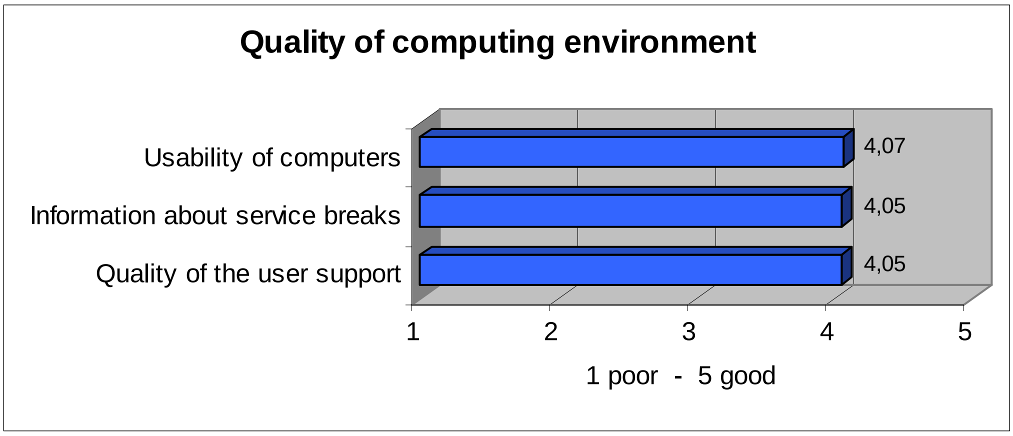

Figure 3.6 Quality of the current computing environment

The interactive usage response time is considered reasonable. Batch job runtimes are a bit too long (Figure 3.7).

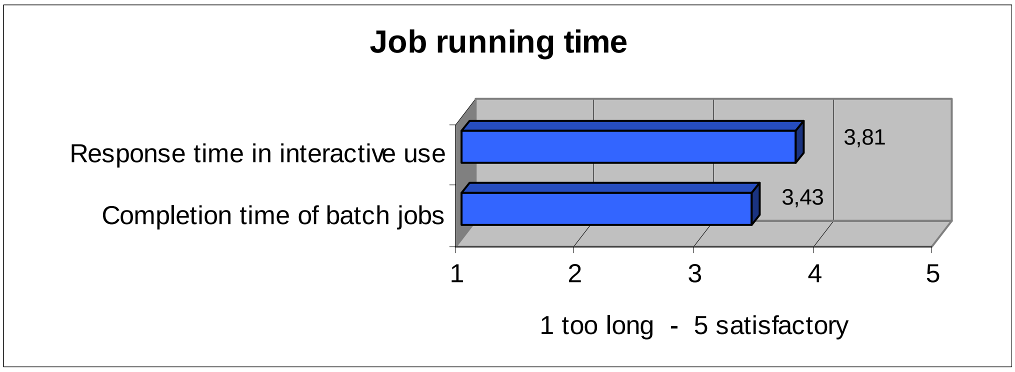

Figure 3.7 Job running time

Figure 3.8 shows how the participants use their computing environment: for parallel and serial computing. Since the users could have both these requirements, they have been asked to specify which fraction of their computing jobs are parallel, serial or other. 30% of participants run entirely parallel jobs, and 11% of participants run entirely serial jobs. Parallel computing is used more often than serial computing. Grid usage and development were also mentioned.

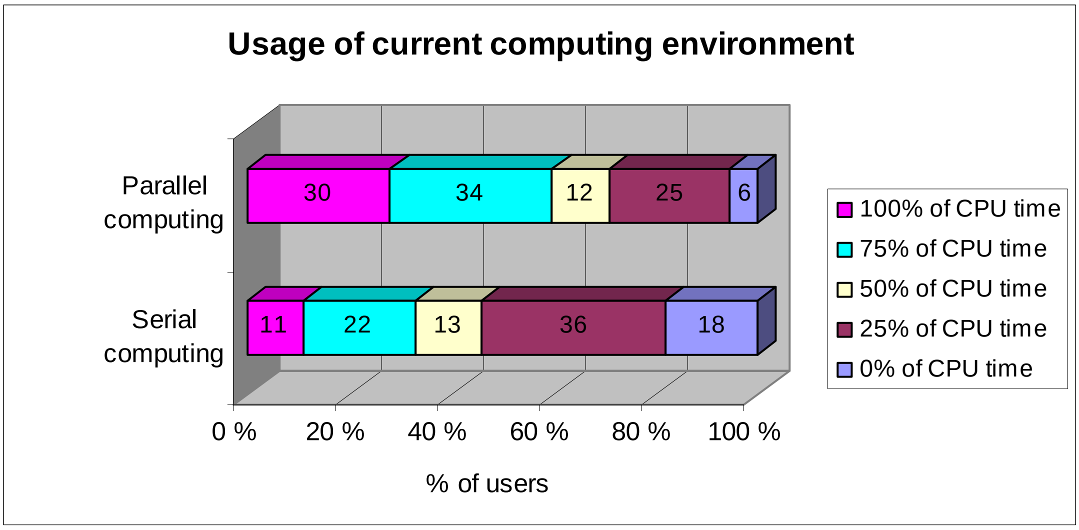

Figure 3.8 Parallel and serial computing

Figure 3.9 shows how the participants prioritise the various features of computational resources. The speed of a single CPU is the most important factor; also the communication between CPUs is important. With parallel codes, scaling up to 8 and 32 CPUs seem to be satisfactory, but there are also needs to scale up to 512 CPUs.

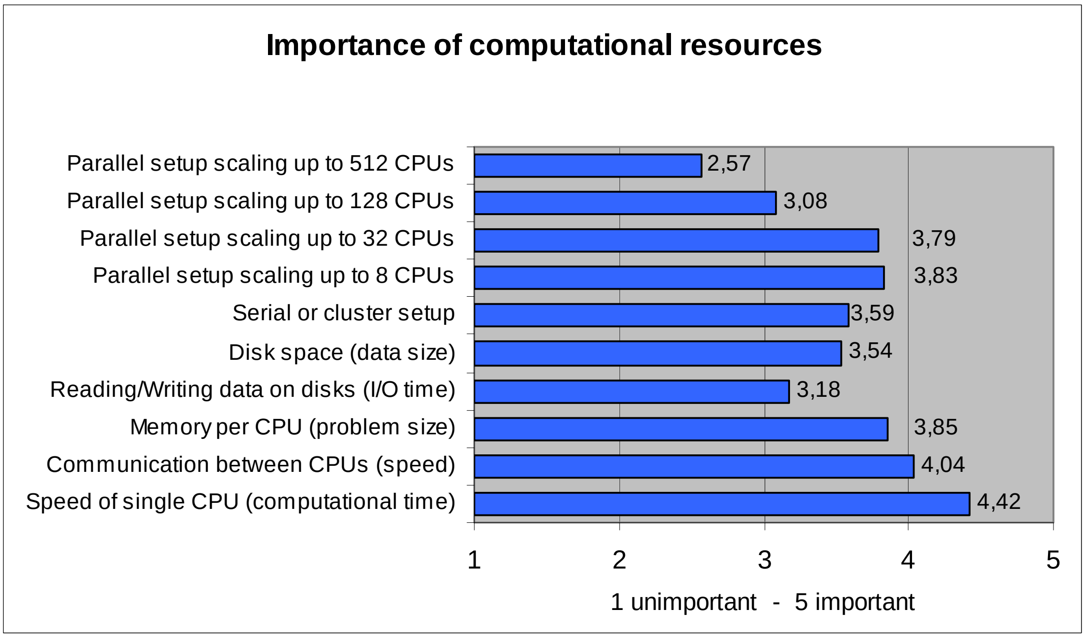

Figure 3.9 Prioritising of the computational resources

For those who prioritised parallel setup scaling up to 128 CPUs, i.e. answered 4 or 5, also fast communication between CPUs is important; 77% of them answered 4 or 5. Those who prioritised parallel setup scaling up to 512 CPUs, i.e. answered 4 or 5, prioritised fast communication between CPU; 85% of them answered 4 or 5.

Figure 3.10 shows the bottlenecks of the participants' current computing environment. You can notice that the speed of a single CPU is the most common bottleneck. Referring to Figure 3.9, it is also considered as the most important factor.

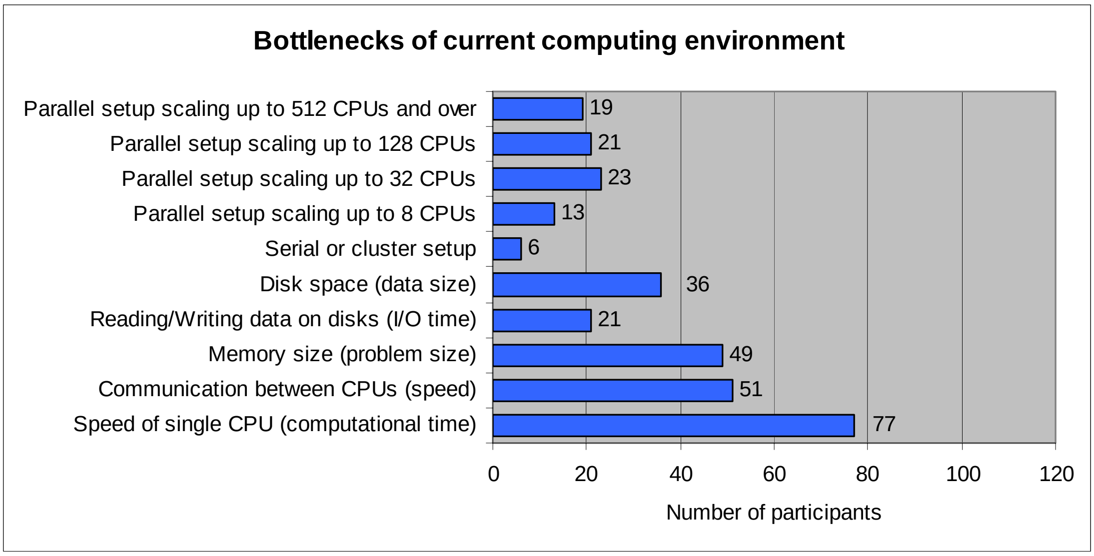

Figure 3.10 Bottlenecks of current computing environment

3.2 Users' expectations by the year 2007

The participants answered freely how they thought their needs would change by the year 2007 with respect to their computing environment. A summary of the answers follows:

- Applications will become more parallel
- Possibility to work with larger problems than today. This is feasible both by an effective use of more processors or an increase in the speed of a single CPU
- Models will become more complex so communications speed between CPUs will become critical
- Memory and storage requirements will increase; they may double or even triple
- Increased need for visualization tools and Grid-enabled analysis tools
- Larger use of cluster setups and Grids
- Some participants think that the changes will not be very big by 2007

4 Application Needs (Questionnaire Part 3)

This section is designed to investigate the applications that the user typically adopted in his/her research activity. The first questions are related to the usage of the computing systems (CPU intensive vs. data intensive), the software programming language (Fortran, C, C++ etc.), the libraries and the compilers required by the application programs (questions 1 to 6). Then a couple of questions analyze the way data are stored and managed (specific standard formats, database usage, questions 7 and 8). Questions 9 and 10 aim at finding out which are the most important tasks in the user's HPC work (code development/porting, data management/analysis etc.), if his/her application can run on parallel/distributed architectures and, if so, which is the adopted parallel computing paradigm. The last question investigates if the user would be willing to spend some time to port his/her application on a distributed environment.

4.1 Analysis of the Results

We first discriminate between two wide categories of applications. The first is represented by "CPU intensive" applications, which requires high performance computing platforms, with powerful CPUs, large memories and high speed mass storage (e.g. fibre channel disks). The second corresponds to "Data intensive" applications, which requires large and fast storage devices, high bandwidth networks and efficient filesystems and/or database software. Since the user could have both these requirements, we have asked him/her to specify which fraction of his/her job is CPU or Data intensive.

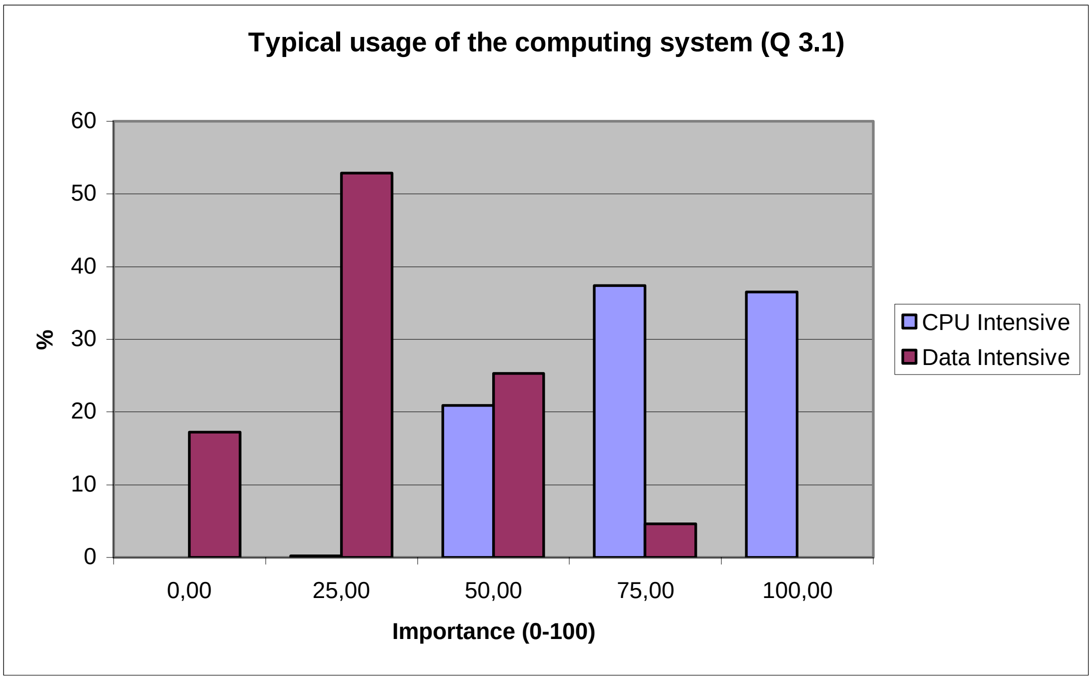

Figure 4.1 Importance of the CPU and Data intensive applications in terms of the percentage of working time spent for each of the two phases

Figure 4.1 shows that computational intensive applications represent a large part of the computing effort for a large fraction of the users, while data handling-management is still not too demanding. This is a trend which is confirmed also by the other questions in this part, and it can be partially interpreted as a bias due to the selected interviewed community, mostly composed by scientists who base their research on numerical simulations. However it must be noticed that data related issues represent a growing concern for researchers. In fact about 25% of the users associated these as 50% time consuming applications.

The next questions analyse in much deeper detail the characteristics of the HPC applications adopted by the users in their work. Question 2 asked if the software is commercial, freeware, open source or self made. Also in this case, since different types of software can be used by the same user, we asked to specify the fractions of the software which belong to each category.

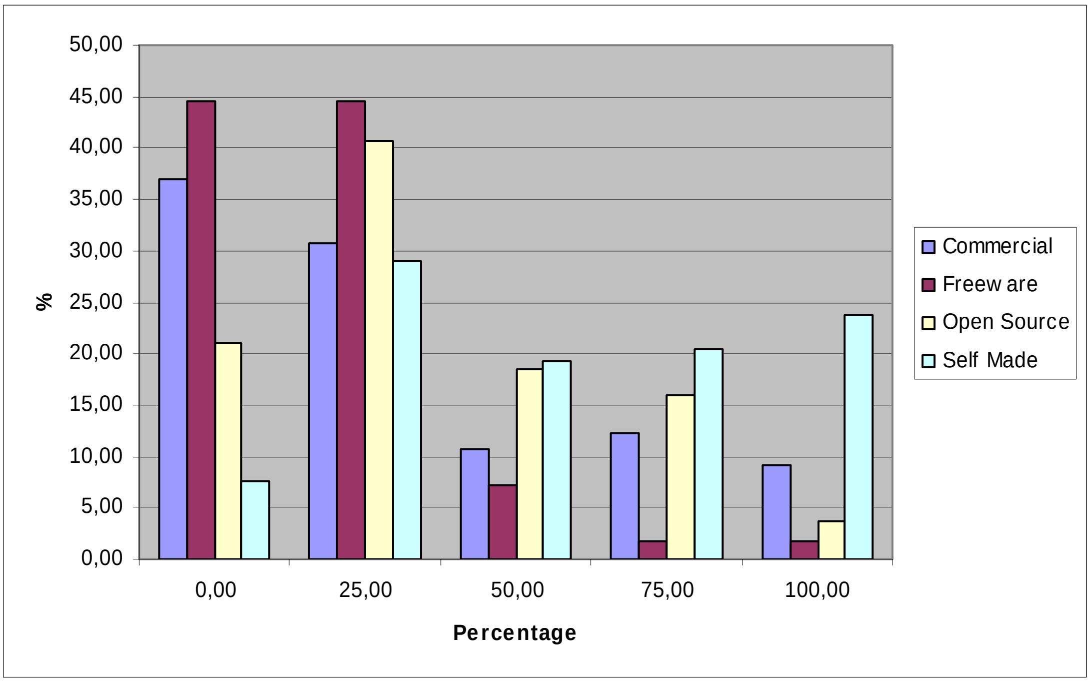

Figure 4.2 Percentage of use of different types of software

Due also in this case to the particular community we refer to, commercial and freeware applications are not so popular. Open source applications are gaining larger credits. But self made software still represents the most typical choice for researchers: almost 25% of the users adopt exclusively self made code. This is due to various reasons: the specificity of the research subjects, little confidence toward third party products, availability of old, well known codes, etc.

We have then investigated the programming language in which user's applications are written. It is not surprising that Fortran and Fortran 90 are the most common languages and are adopted by about 47% of the users. However it is interesting to notice that C and C++ codes are more and more diffused. More than 40% of the scientific programmers use them. The rest of the applications are written in Java or interpreted languages like Perl, Matlab, IDL. The codes are compiled using a large variety of compilers. HPC users tend to use vendors (IBM, SGI, SUN…) proprietary compilers, since they can provide a high level of optimization and the best performances. However the diffusion of cluster architectures (e.g. Beowulf and Linux clusters) has led to an increasing usage of more general products like the GNU, the Portland Group and the Intel compilers. Figure 4.3 summarizes these results.

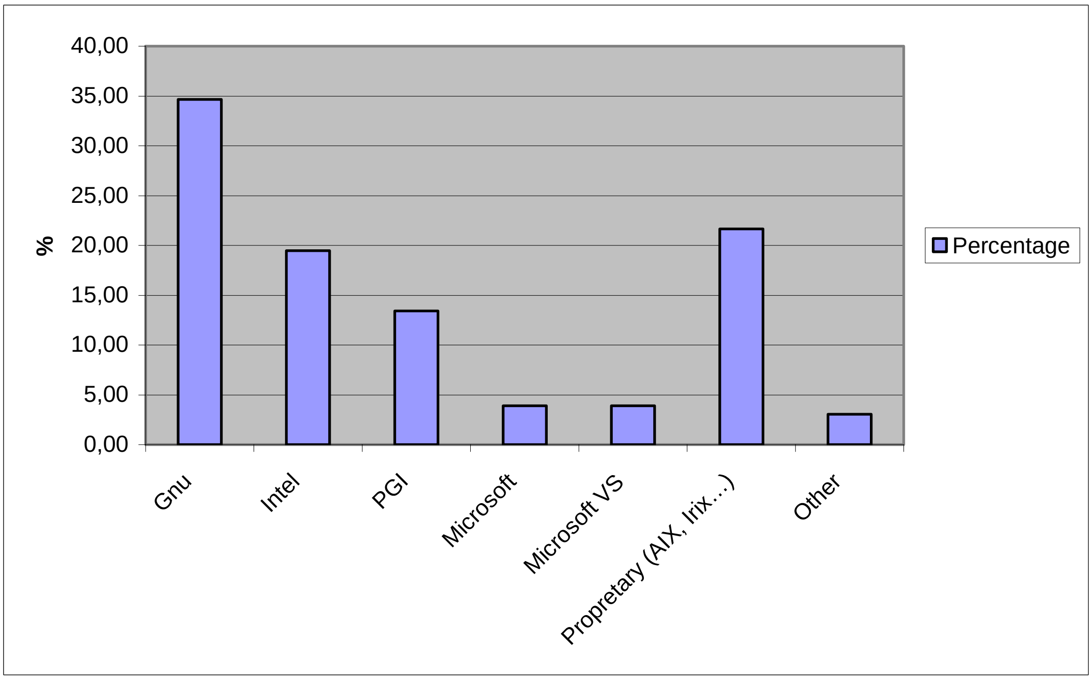

Figure 4.3 Diffusion of different kinds of compilers

The analysis of the codes details is completed checking if specific libraries are used and, if so, which libraries are mostly used.

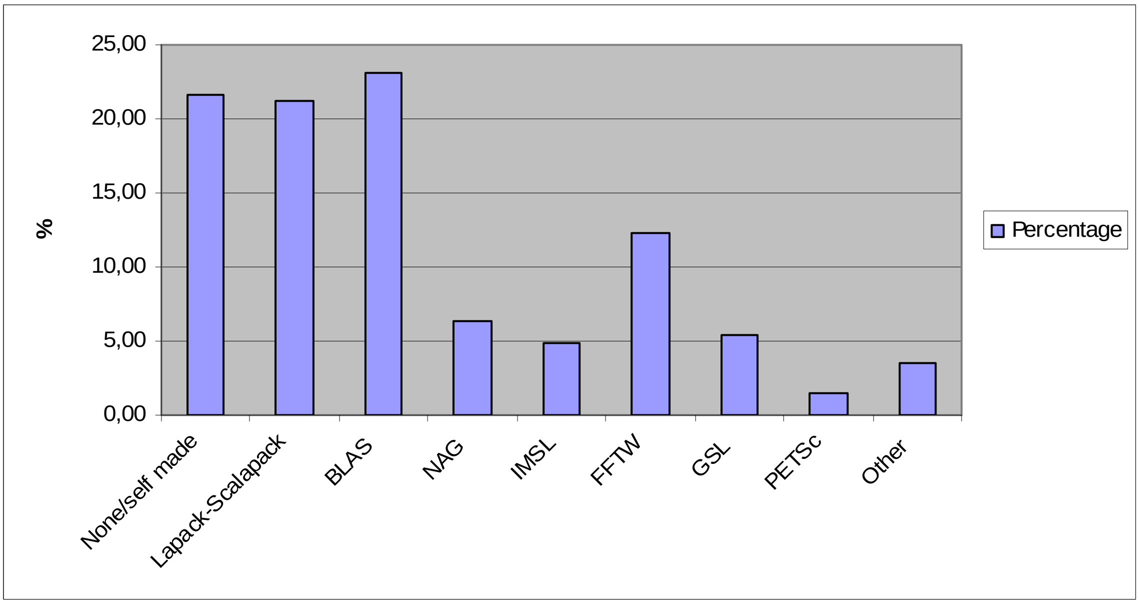

Figure 4.4 Usage of scientific libraries in users' codes

The results are shown in Figure 4.4. It is interesting to notice that, again, a large fraction of the community (more than 20%) does not use any available library, building completely by itself the codes. Furthermore, open source tools, like FFTW, GSL and PETSc are getting a growing interest from the developers, since they are able to provide performance and algorithmic quality comparable to commercial products.

We have generalized the analysis of available software tools to the software development environment, ranking the quality of available editors, compilers, scientific libraries, performance analysis tool and debugger, with marks that range from 1 to 5. It is not surprising that editors are the tools with the highest ranks: more than 85% is between 4 and 5 and none got 1. However, also the score of compilers and libraries is high. About 84% of the compilers and 67% of the libraries were ranked between 4 and 5. The situation is different for profilers and debuggers. In this case most of the available tools are ranked around 3, but a fraction of the users (between 5 and 9%) ranked them with the lowest mark. This points out that these tools in the HPC framework, where their availability is critical, are not satisfactory and more effective instruments would be necessary.

At this point we have moved to data and their management. First of all, we have analysed which are the most common data formats in the scientific community. The outcome is presented in Figure 4.5.

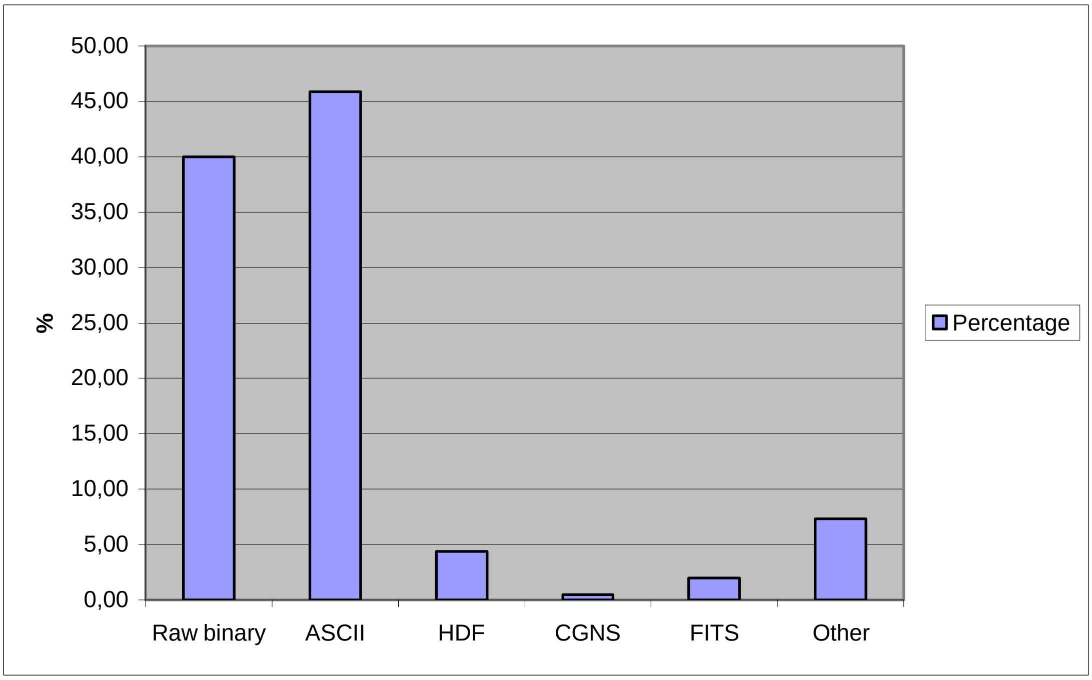

Figure 4.5 Data formats in the scientific community

Most of the users does not adopt a standard data-format but handle the data in the raw or even ASCII formats. This prevents data exchange, federation, re-usability and strongly limits the collaboration and the diffusion of knowledge. Then, we have asked how data are managed and the results point out that data are considered a minor issue by the scientific community. In fact only 25% of the users manage its data by advanced database tools (Oracle, DB2, MySQL). This can be partially justified observing that typical relational databases cannot handle effectively large data-sets of raw data (typical outcomes of numerical simulations). Therefore, also a strong effort by the database developers would be required in order to overcome this difficulty. However it is meaningful to stress how only a low fraction of the users, 77 over 125, has answered to this question, further emphasizing how presently the subject is not considered relevant.

The investigation of the involvement of the users in each part of the "numerical pipeline", from the code development, to the run, to data analysis and visualization, is the next step of our study. Each different aspect is ranked between 1 and 5. The average mark and the related standard deviation (sigma) are presented in Table 4.1. If the standard deviation is over 1.00 the opinions can be considered to be divided, between 0.80–1.00 unanimous, and under 0.80 very unanimous.

Table 4.1.

Mark	Code development	Code porting	Optimization	Run	Data management	Data release	Data analysis	Visualization
Avg. Mark	4.07	3.32	3.87	4,21	2.87	2.72	3.61	3.58
Sigma	1.15	1.35	1.22	1.04	1.18	1.26	1.28	1.19

Most of the users gave high scores to the first four phases, which are all strictly related to the code development and run phases. Again, less attention is given to the data, with little interest in data management and sharing of the results. There is a certain interest in the data visualisation step, which is probably seen as the most direct way of analysing the results and mining the information contained therein.

Finally, we have focused on parallel applications and the possibility of exploiting distributed architecture and the Grid environment. First we have asked the users if their applications are already parallel (if not, if it could be useful to parallelise them) and, in this case, what is the paradigm used for the parallelisation. Figure 4.6 summarises the answers.

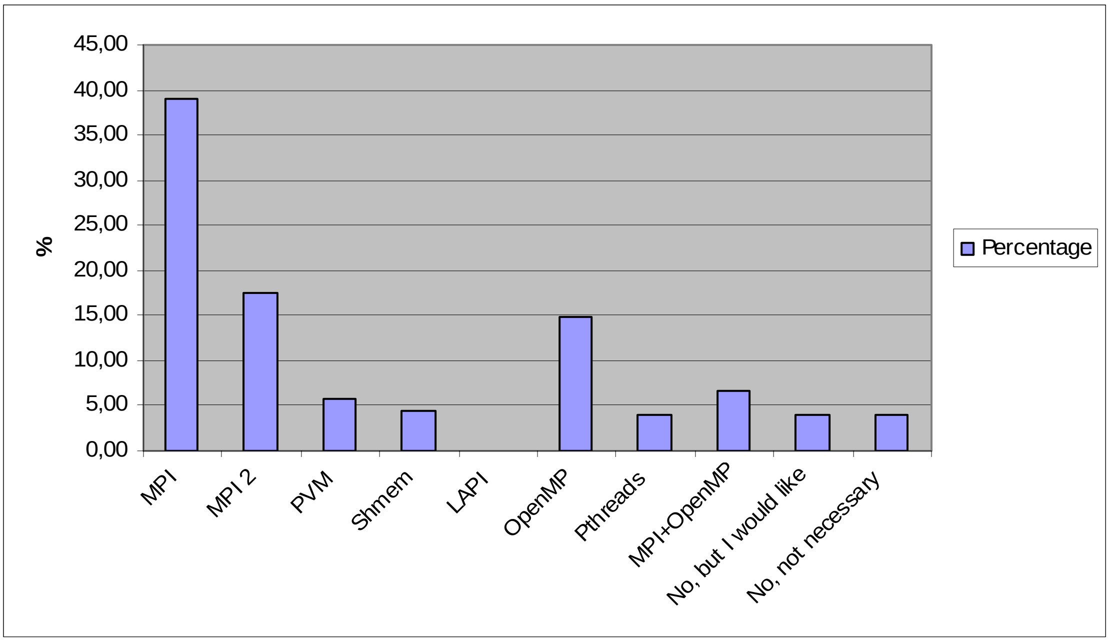

Figure 4.6 Paradigms of parallel computing adopted by users' applications

The most common paradigm is the Message Passing implemented using the standard MPI-MPI 2 libraries. More than half of the users adopt MPI in order to make the code efficient, scalable and easily portable. It is quite surprising the large diffusion of MPI 2, since this is not yet a completely stable and reliable product. Some older codes still adopt the PVM library. The diffusion of the shared memory paradigm, based on OpenMP or P-threads, is seriously limited by its architectural requirements (shared memory is strictly necessary) and low scalability typically above 16 processors. Proprietary products, like Shmem (CRAY-SGI) and LAPI (IBM) are not widely used due to the low portability of the resulting code. Also mixed codes (MPI+OpenMPI) are not common due to the complexity of such technique.

We have then verified the interest of the user toward the Grid infrastructure. In particular, the user is asked if he/she is willing to spend some time in porting his/her application on the Grid. Figure 4.7 presents the statistics of the answers.

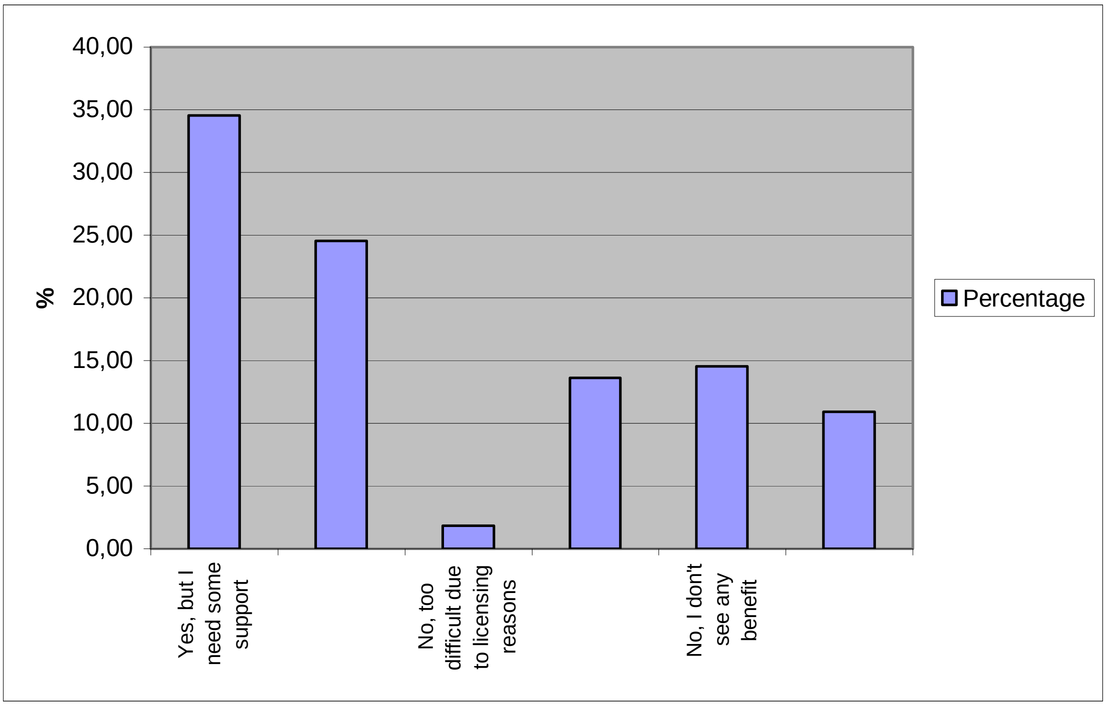

Figure 4.7 Users feeling towards the Grid

The results indicate that a large fraction of the users is interested in making use of the Grid infrastructure, even though it is still seen as an unfriendly environment that requires help from experts to deal with. However, more than the 20% of the users are not attracted by the Grid or do not see any benefit in using it. Finally, it is interesting to notice that an important fraction of the users, more than the 10%, can already exploit the Grid infrastructure for his/her scientific work.

5 Grid Environment Experience (Questionnaire Part 4)

The last part of the questionnaire focuses on the Grid technologies. The main object is that of investigating the familiarity of the user with respect to the Grid environment. In particular, the first half of this part is dedicated to analyse the user's experience with Grid based applications, if Grid represents a completely new subject, if he/she has already used Grid based tools or even if he/she is a developer of such instruments. In the second half, we investigated the user interest in "entering the Grid" sharing computational resources, software, date and, generally speaking, knowledge.

5.1 Analysis of the Results

First we analysed the general expectation of users toward the Grid environment. We have asked the user to rank (five degrees, from "non important" to "important") the importance for their work of the following aspects:

- Have higher computing power available. Most of the users (more than 85%) consider this aspect the most important, ranking it between 4 and 5. This is a clear consequence of dealing with the HPC users' community.
- Have a single point of access to HPC resources. This means submit a job and let the system find out, proper, tightly coupled computing resources where it can run. This question got a flat ranks distribution, indicating either this is not an interesting functionality for the users or that its importance is not yet clear.
- Store data in distributed databases. More than 60% of the users consider this aspect of little importance, ranking it between 1 and 2. Only for 8% of the survey the mark is 5.
- Have a virtual collaboration environment to work with "remote" colleagues. In this case the users are equally divided between those who consider important (35%) and those who consider useless (40%) this aspect. Probably this is due to a lack of experience with such tools and also to the fact that distributed workshops, which are one of the most important and known aspects of collaborative work, were not explicitly mentioned in the question.
- Distance learning and training. More than 63% of the users ranked it as "non important" (marks 1 and 2). Only for 7% of the users this item can get the highest score.

Then we have examined the concrete experience of the users with Grid based applications, asking if the user has ever attended a web-workshop or other distributed collaborative working session and if the user has ever used some emerging Grid tools like Access Grid, Virtual Observatory and Cactus. In both the cases, a large majority of the users had never got in touch with one of these tools. Only about 15-20% of the users experienced them. However, most of the people are interested in starting to deal with such tools. This is further confirmed by the fact that most of the users have not even heard about the most popular Grid middleware, like Globus or Unicore and, in any case, they had very little experience with them. Figure 5.1 shows the statistics. This can be only partially explained with the fact that we are dealing with a scientific community, while these packages refer principally to a technical context.

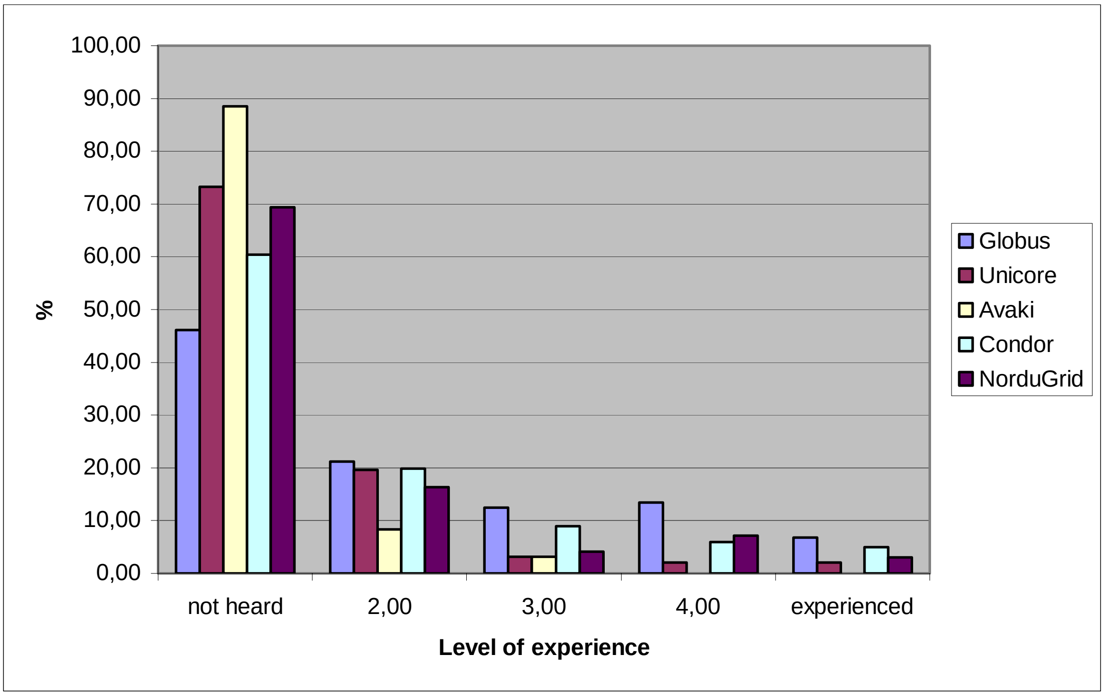

Figure 5.1 Users' experience about the Grid middleware

However, although the majority of HPC users have not directly experienced the Grid, they have clear ideas about the basic features it should have, as emerges from the analysis of Table 5.1. The table shows the average mark (in the range 1-5) associated to different features of the Grid.

Table 5.1. Which of these features do you think are critical for a Grid based distributed system?

Feature	**Avg. Mark**	**Sigma**
Web access	3.27	1.21
Single image Filesystem	3.42	1.13
Security	3.90	1.15
High network bandwidth	4.30	0.98
Schedulers	4.12	0.86
Large aggregate computing power	4.32	0.82
Large aggregate data storage capacity	3.65	1.14

High performance (network bandwidth and computing power) and efficiency (schedulers) represent the critical aspects for most of the users. Security is another important issue. Web portals and single points of access and single image filesystems are instead considered desiderata, but not crucial points. Large data storage capacity divides the community between those who consider it a basic requirement (about 32%) and those who see it as a useful but not indispensable aspect (rank 3, 33%).

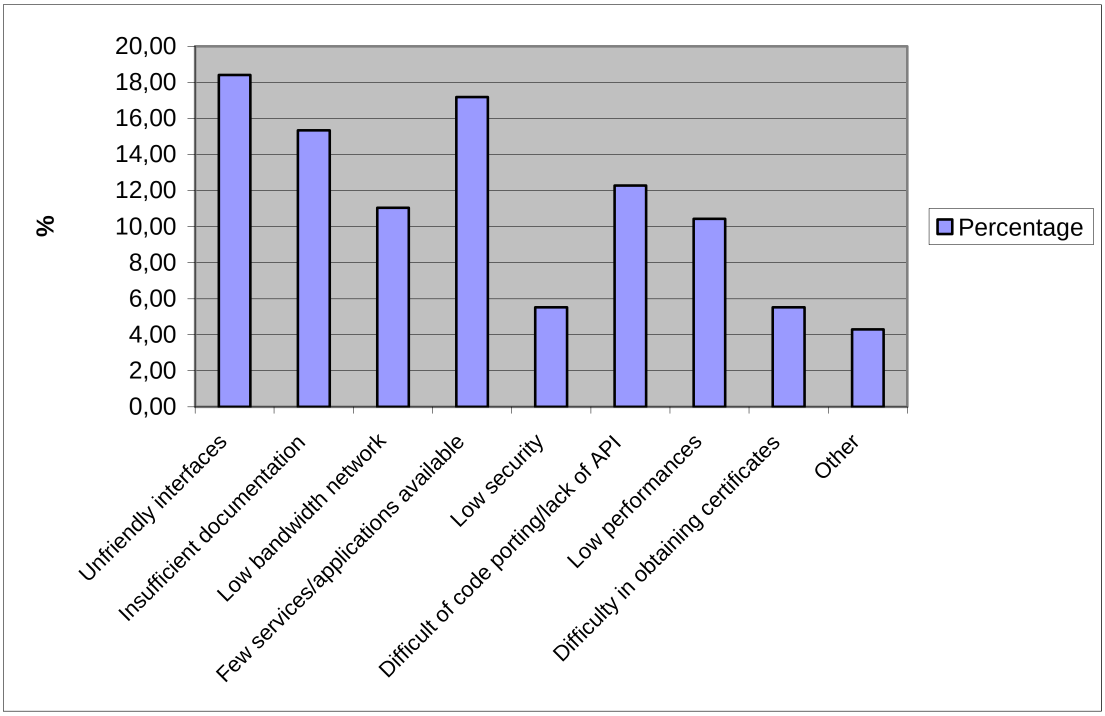

Figure 5.2 Main difficulties evidenced in the present usage of the Grid

The main difficulties found by those who have already experienced the Grid environment represent our next concern. The statistics presented in Figure 5.2 shows that, at present, the most critical aspects are not related to the hardware or the performances, but to the difficulties of interacting with such framework. Unfriendly interfaces, lack of documentation, insufficient services make it the approach to the Grid difficult and tricky. This, however, is an unavoidable consequence of the development stage in which the Grid infrastructure presently is.

The first half of this part of the questionnaire closes asking if any of the users is involved in projects of Grid development. It came out that more than 27% is involved in such projects, which is not a negligible fraction of the community, further emphasising the interest of the scientists for this large scale infrastructure.

The second half, investigates if the user is interested in taking part to the Grid development process and in particular in the sharing and federation of tools, data and knowledge.

We have asked the users which of the proposed services, if any, could be hosted by their institution in future. The services are: HPC resources, data storage and collaborative work infrastructures. Figure 5.3 shows the results. The user could also answer indicating either that he/she gets the computing services from other institution (outsourcing) or that he/she cannot host any kind of device. It is encouraging that more than 80% of the users is willing to contribute in some ways to the Grid infrastructure.

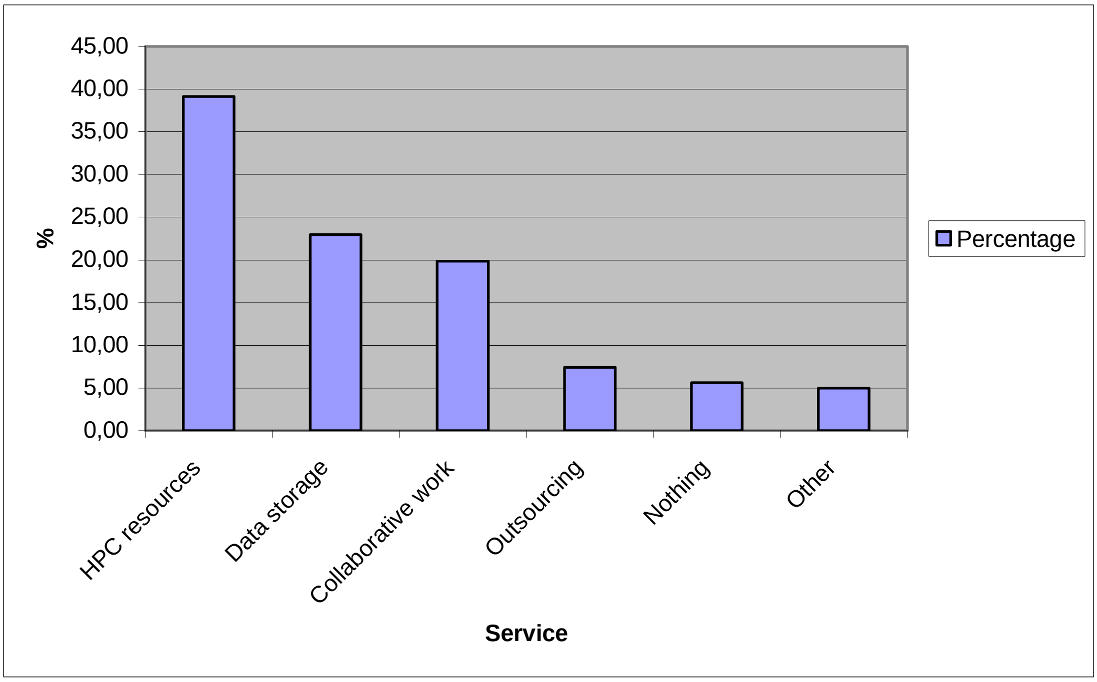

Figure 5.3 Users that are interested in hosting some kind of Grid enabled devices

Then we have asked the user if he/she is willing/interested to make his/her code and data available to the scientific community to be used on the Grid.

75% of the users are willing to release the code to the community (even if most of them need a preliminary porting procedure). 12% use licensed software, while only 13% of the users do not want to share his/her numerical codes. The result for data is quite different. In fact only 30% of the users would give unconditionally his/her results to the community. 37% would release data only under specific conditions (e.g. acknowledgements, grants etc.). But there is also 33% who are not willing to make the data available to the community. This for many different reasons: results are already available on journal and web sites, data are too problem specific, they can be asked directly to the author and so on. The property of the data comes out to be a critical issue for scientists, even if more than 70% of the users think that sharing of information and knowledge could be an important way for improving the visibility and the popularity of their work.

5.2 Users Comments

At the end of the questionnaire we have asked the users to propose any idea, suggestion about HPC, Grid and their possible developments. In the following we report the most interesting answers:

- Try and explain exactly what the Grid is and how it works, me for one have been hearing a lot about "the Grid" but very little about what it actually does or can do.
- I would strongly appreciate more API documentation. I think it is quite poor and difficult to follow.

- The sooner we settle on middleware standardization the better. Right now small research groups such as ours only gather experiences but unfortunately spent a lot of effort following dead-ends because of rapid (and rush?) changes of directions (gt2 to OGSA to Web Services). For example, a recently accepted conference paper using OGSA became outdated even before the camera ready date.
- It is important that HPC and Grid resources are easy to access. This means that batch queue systems and computing environment ought to be rather similar. My experience this far is that each HPC system has its own working environment, and the time needed to learn the important features is relatively long. This makes it difficult to utilize several HPC sites at the same time. It is also difficult to get an overview of the output files when these appear at several systems at the same time. Some solutions for the logon process are quite complicated and I definitely prefer systems that use SSH. This also makes file transfer fairly simple. It is also important that common software for important applications is made available in a standardized manner. It is also important to work out good solutions with regard to licensing issues. Some computer codes tend to be very expensive and difficult to port to computer environments beyond the home institution.
- I think the most important issue for a Grid is the efficient queuing system which is a global queuing system (if this is at all possible).
- The Grid should be available for people running other platforms than Linux, for example MacOSX.
- Make it as much transparent as possible. If the Grid will offer more resources than what available at local computing centres, than a human resources can be used for that. If not, it is not worthwhile. If yes, this must be evident.
- The benefit of having a single interface to a huge number of resources is the main advantage for me.
- Providing lots of storage (to the point that everything is online) is important. Need to worry about latency sensitive computations.
- You might want to implement the largest program packages in the Grid environment before people start to get accounts.

6 In-Depth Interviews

6.1 Method

Seven in-depth interview questions were prepared to representatives of significant computational science research groups or organisations in Europe to solicit their views and opinions on future requirements for HPC. Altogether 11 in-depth interviews were conducted. This activity was led by CINECA and CSC, with assistance from other ENACTS participants, who selected the representatives and conducted the interviews.

The interview questions were sent to the representatives beforehand. The representatives had a possibility to decide on the interview method; half of the interviews were made face-to-face and the other half remotely by phone or email or a combination of these two. The interviews were conducted in autumn 2004.

6.2 In-depth Interview Questions

The following questions were asked during in-depth interviews as part of the ENACTS Survey of Users' Needs:

1. What do you think will be the future HPC requirements of the scientific community? More CPU power? More storage capacity? Data management tools? Data analysis/visualization tools?

2. Which do you think would be the future “winning” HPC architecture for the scientific community? Shared memory systems, parallel machines, low cost clusters (Beowulf clusters), distributed systems, or something else? Will the researchers from various fields have differing needs?

3. Both HPC applications and experiments/observations are producing huge amounts of data that have to be stored, managed, mined, and analysed. How do you think the scientific community will face this problem? Which are the most important instruments to deal with such issue?

4. The international community is producing a big effort in the development of the Grid infrastructure. What is your feeling with respect to this new environment (present status, future developments, technological trends etc.) and how do you think the Grid can be useful for the scientific community in your country? What areas of research will benefit the most from future systems?

5. Do you think that the scientific community is willing to accept the sharing of knowledge, resources, tools and data, as proposed by the Grid idea? How do you think issues like knowledge property, data access etc. can be faced?

6. Do you think that virtual organisations, distributed workshops (e.g. Access Grid), or distance learning can be useful tools for the development of science and, more generally, of knowledge? What are the benefits? What instruments would they require to be more and more effective?

7. In your opinion, which are the most important requirements and expectations of the scientific community with respect to the near future technological development? How do you think the role of HPC centres will change?

6.3 Interviews

In the following sections we present the collected in-depth interviews. They are presented in alphabetical order, with respect to the family name of the interviewed.

6.3.1 Interview with Axel Berg, SARA Amsterdam

Dr. Axel Berg is the head of the HPC department of the Netherland's Stichting Academisch Rekencentrum Amsterdam (SARA).

6.3.1.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Dr. Berg can be summarised as follows:

1. Both CPU power and data storage are critical. However the latter is becoming more and more important.

2. The winning HPC architecture will therefore be a combination of HPC architectures that serves the variety of applications of the scientific community in the most cost-effective manner.

3. Formation of communities, sharing resources within and outside the community and the adaptation of tiering models are crucial for a full exploitation and management of the data.

4. Many of the current Grid infrastructures are still experimental. A strong focus should be put on production quality of service, multi-disciplinary Grids, usability and dissemination to turn around the Grid developments to real user benefit.

5. The scientific community is very willing to accept the sharing of knowledge, resources, tools and data. Property, secure access, accounting and scheduling are crucial issues that have to be resolved also on a political level.

6. Tools for distance collaboration are very useful but can never fully replace life human interaction.

7. The most important requirements and expectations are the provision of the adequate and usable integrated resources and expert support. Some HPC centres could become more Grid service and support oriented rather than Grid resource oriented.

6.3.1.2 Interview

1. CPU-power: there has been and will be a nearly unlimited need for CPU power in the scientific community. This community is continuously defining new scientific challenges that can only be attacked by the computing power that will be available tomorrow. However, simply adding more and more computing power is a too simple solution here, much emphasis and effort should be carefully put in tuning and mapping hardware and software architectures to real applications, and for instance in performance modelling, to ensure cost effectiveness in the future, the risk of widening the gap between peak-performance and application performance is large.

Storage capacity: requirements are increasing rapidly here at exponential speed (e.g. at SARA the amount of data stored has doubled for four consecutive years). This is due both to the increase of data produced by complex simulation and to data produced by large scientific instruments (in life sciences: DNA-sequencers, micro-array facilities; astronomy: radio telescopes, medicine: HR-scanners; high-energy physics: particle accelerators; etc. etc.). In practice today the amount of data being stored is usually budget limited, following Parkinson's Law: “Data expands to fill the space available for storage.“

Data management tools and also data modelling tools are critical to cope with the huge amounts of experimental data that need to be analysed to transform data into knowledge.

As experimental data grow exponentially, new efficient scientific visualization methods are required to analyse data the large amounts of data. HPC centres should develop and provide expertise in this field as this is becoming a bottleneck in data analysis by the scientific community. In general data storage and management has become more important than CPU-power.

2. There will be no winning single HPC architecture; the challenge is to map carefully the applications and load of various scientific disciplines onto a variety of computing architectures. This challenge is becoming larger and easier at the same time due to the development of (distributed) Grid infrastructures. Larger in the sense that the 'transparent' and 'blind' submission of jobs to a Grid infrastructure will enlarge the amount of jobs that will be run semi-optimally and less well-tuned to a particular system in the Grid infrastructure. Easier in the sense that the total pool and variety of HPC resources for a particular community is potentially enlarged, increasing in principle the opportunity of finding a HPC system that is well-tuned for a particular application.

The winning HPC architecture will therefore be a combination of HPC architectures that serves the variety of applications of the scientific community in the most cost-effective manner. Usability is another key issue to optimize the total costs of application performance, and also to ensure a winning strategy towards new scientific disciplines to use HPC systems (e.g. life sciences). Like in the past, researchers from various scientific fields will have different needs in the future.

3. The communities that face extreme data challenges will face these problems differently. Communities that traditionally are performing data intensive sciences (e.g.

HEP community) form usually a strong world-wide community and they are preparing themselves well in terms of technology, data distribution and funding models. Scientific disciplines that are becoming data intensive (like life sciences, cognitive sciences, astronomy, etc.) should learn from the communities that traditionally are data-intensive. Formation of communities, sharing resources within and outside the community and the adaptation of tiering models are important here.

4. Many of the current Grid infrastructures are still experimental and not ready to use as a production facility by the scientific community. Only the last 1-2 years production Grid infrastructures have emerged, and this process will (have to) continue in order to survive. For a successful future of Grid infrastructures, strong focus should be put on production quality of service, multi-disciplinary Grids (one should not build general purpose Grids!), usability and dissemination to turn around the Grid developments from strong technology push to application pull and real user/application benefit. There should be as least as much focus on the adaptation of the Grid infrastructures to the user requirements and skills as there is focus on adaptation of the users/applications toward these new Grid infrastructures.

Grid infrastructures will be very useful for the Dutch research community as it enables sharing resources and data among various scientific communities inside and outside the Netherlands. Besides the HEP community in the Netherlands (collab. in EGEE etc.) other communities like earth observation, climate modelling and astronomy are using a Grid infrastructure today, and others like e.g. life sciences, food informatics and medical sciences will be using a Grid infrastructure in the very near future for their day-to-day research.

5. The scientific community is very willing to accept the sharing of knowledge, resources, tools and data as long as this sharing is limited to their own scientific disciplinary community and within existing scientific collaborations. In particular peer-to-peer sharing between research groups will be highly valuable and appreciated. Besides technology challenges in areas like knowledge property, secure access, accounting and scheduling, these issues have also many aspects and challenges that need to be resolved on a political level. The real challenge here is joint communication on both the political and technical level to design well-working solutions and mechanisms.

6. Tools for distance collaboration are very useful but can never fully replace life human interaction, which is essential for good collaboration and good science. Benefits are mainly efficiency (e.g. in terms of travelling). To become more effective today these instruments should become more commodity first, today technology is still difficult and requires usually a set of non-standard equipment and software and a special skills to operate and use it.

7. The most important requirements and expectations are simply the provision of the adequate and usable integrated resources and expert support. The role of HPC centres could change in the sense that due to integrated (European) e-Infrastructures resources and expertise will be more spread within and over countries and that therefore some HPC centres could become more (Grid) service and support oriented rather than (Grid) resource oriented.

6.3.2 Interview with Luca Biferale, INFM Rome

Prof. Luca Biferale works at the Physics Department of the University of Rome “Tor Vergata”. His research activity deals with Turbulence, Turbulent Transport, Chaotic Systems, Stochastic Processes, Multifractal, Wavelets, Monte Carlo Methods, Numerical Simulations, Pseudo Spectral Methods, Lattice Boltzmann Equations, Data Analysis, Phase Transitions, Renormalization Group.

6.3.2.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Prof. Biferale can be summarised as follows:

1. Different research fields require different systems.

2. Independently from architectures flexible and effective access policies to HPC facilities are necessary.

3. Storage facilities must develop together with HPC architectures and the data must be easily accessible.

4. Grid infrastructure and Grid computing do not still exist. Grid must go beyond the actual status of "meta job scheduler" and deliver a real complete service.

5. There are no particular problems in sharing knowledge and resources.

6. Promising, but I have never used such tools.

7. HPC centres must become the backbone of a true Grid infrastructure with the aim to allow easy, sound and permanent access to the scientific community using HPC.

6.3.2.2 Interview

1. My impression is that there is not a unique strategy. Different research fields would require different developments to see their necessities optimised. Our research, which mainly concerns Direct Numerical Simulations of fluid-dynamics, is demanding computation on parallel architecture with very intense communication requirements.

2. Different fields require different architectures. The only important requirement which I think is shared by the whole community is the policy of access to the HPC facilities. We need flexible and transparent procedures to select and manage the projects.

3. Massive storage facilities must go together with HPC facilities. Data must be easy accessible, web based facilities to sample and download (part of) huge data set would be extremely useful.

4. My feeling is that Grid infrastructure and Grid computing do not still exist. At least concerning their ability to deliver anything that has added benefits with respect to standard HPC centres. Most of the intense computing applications require high

communication performance between nodes. I am not aware of any successful research project based on Grid. The only exception up to now is data post-processing. Grid must go beyond the actual status of "meta job scheduler" if we want to give a boost to the HPC research community.

5. I am not an expert of this. I do not see any big problem from the side of the scientific community.

6. I am not an expert of this. It looks certainly useful, professionally I have never used this tools.

7. I think that there are two different expectations. First, the obvious demand for technological improvements especially concerning the inter-node communication. Second, even more important, is the flexibility and transparency in the access to the HPC tools. Scientists need to be able to access them on a daily base, without barriers imposed by security requirements and/or knowledge property issues.

HPC centres must tighten their links, both physically and scientifically. They must become the backbone of a true Grid structure with the aim to allow easy, sound and permanent access to the scientific community using HPC.

6.3.3 Interview with Gianluigi Bodo, Astronomical Observatory of Turin

Professor Gianluigi Bodo is an associate astronomer at the Astronomical Observatory of Turin. He is presently teaching “Computational astrophysics” and “Cosmic physics” at the University of Turin. His main research interests are related to the physics of extragalactic jets and numerical simulations of jets formation, evolution and properties.

6.3.3.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Professor Gianluigi Bodo can be summarised as follows:

1. High performance data management, visualization and analysis tools will be crucial.

2. Beowulf can represent a winning solution. However, high end systems are still necessary.

3. Specific instruments for data handling must be developed in order to face the problems related with their increasing size.

4. The Grid is not crucial for HPC. It is much more important for data sharing and access to distributed database as in the case of the Virtual Observatory in astrophysics.

5. Resources sharing is essential but tools to support it must be developed and rules and protocols must be defined.

6. Virtual organization and workshop can be useful but direct interaction is always important.

7. HPC center must provide computing and data services, but also support and scientific collaborations.

6.3.3.2 Interview

Gianluigi Bodo, Professor at the Astronomical Observatory of Turin, gave the following answers to the ENACTS in-depth interview questions:

1. The CPU power of existing HPC systems allows large scale simulations with the production of increasing amounts of data. In my opinion one of the critical points is the handling of such huge amounts of data. Therefore, in my opinion, data management tools and tools for data analysis and visualization are two areas in which improvements are required. Due to the large size of the data all these aspects will require HPC systems. Therefore an important role will be played by the development of platforms where it will be possible to analyse and visualize large dataset, maybe already obtained from simulations.

2. Of course the type of architecture depends on the kind of problem to be considered. Beowulf clusters due to their cost effectiveness will probably be widely used for mid-range simulations. However shared memory system or high end computing platform will be necessary for extremely computational demanding or very specific applications.

3. Sharing of data will become always more and more important. The Grid will play an important role since it will provide instruments to deal with distributed dataset in efficient and effective ways. Furthermore it will minimize the need of data replication and movement over the network. Obviously the technology and in particular the database technology must develop to facilitate and make more and more efficient the management and handling of large amount of binary data. In particular tools of data mining, reduction and decimation could be very effective in optimising the process of data analysis.

4. I Don't think that HPC will get substantial benefits from the Grid infrastructure, except from extremely large projects. In my field of research (astronomy) I think that distributed archives and data analysis can get good benefit from the Grid infrastructure. In particular a big effort is being produced by the international community to build up the “Virtual Observatory” (VO). This effort is coordinated by the International Virtual Observatory Alliance (IVOA), which is a transnational organization and has started a joint effort in order to create a distributed infrastructure (the VO) which could provide the maximum fruibility of data and common tools to deal with them.

5. I think that the sharing of knowledge, resources, tools and data would be to a large extent accepted by the scientific community, if technologically possible. I think that the community wants to share data and tools, but there is the necessity to define a proprietary period, after that the data and tools become free. Furthermore access policy must be formalized.

6. Virtual organizations and distributed workshop can potentially be useful tools. Easing long distance and distributed collaborations will favour the exchange of ideas and the birth of new projects. Furthermore it can be a way to save money and time. However direct interaction between people is always important

7. Networking is one of the most important requirements: it is important to communicate and exchange ideas, tools and data quite fast. I think that HPC centres risk to become places for data storage only, while it is important to have also large supercomputers (not only Beowulf). It will be very important also that computing centres provide not only technological resources but also competencies, expertise and scientific collaboration.

6.3.4 Interview with Roberto Capuzzo Dolcetta, University of Rome "La Sapienza"

Roberto Capuzzo Dolcetta works as associate professor at the Physics department of the University of Rome "La Sapienza" where he teaches "Physics of Gravitation". His main research activities are related to smooth particle hydrodynamics and N-body evolution, applied to the study of multi-phase, self-gravitating astrophysical systems, dynamical evolution of globular cluster systems in elliptical galaxies, and resulting AGNs and identification of OB associations in unresolved galaxies.

6.3.4.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Associate Professor Roberto Capuzzo Dolcetta can be summarised as follows:

1. Computing power is still the central issue. Data must be produced cleverly.

2. Both distributed and shared memory systems should be available to the research community to face at best different problems

3. Data issues should be treated with a clever approach more than with sophisticated tools

4. Grid is still a vague concept. Working prototypes must be shown to convince people to use such infrastructure

5. Knowledge sharing is a benefit but it must be governed by precise and up-to-date rules.

6. Collaborative working is important but specific infrastructures (e.g. large bandwidth networks) and software tools are still required.

7. HPC centres should provide not just computing power but solutions like expertise, support, collaboration etc.

6.3.4.2 Interview

Roberto Capuzzo Dolcetta, Associate Professor at the University of Rome "La

Sapienza", gave the following answers to the ENACTS in-depth interview questions:

1. The specific requirements depend of course on the type of users. Some will need an extension of shared memory computer power and other prefer distributed large Beowulf. I am in the first category.

More CPU power is definitely a basic requirement. An increasing storage capacity is important, but the need of larger and larger storage corresponds to an accumulation of data output which is overwhelming the human capacity to handle with all these data. Top science is often done with clever consideration of modest (in size) outputs.
Data management and visualization tools need of easy and friendly interfaces for that, that are often not available.

2. Shared memory systems are objectively the best being the fastest and the easiest to program. The problem is that present technology does not allow to have enough architectures with enough processors to guarantee highest speed and at the same time good scalability. A good computing center should have a large Beowulf and/or a distributed system based on vectorial CPUs AND a good shared memory machine.
Researchers from various fields can have differing needs if they are on top of their fields. See my comment at point 1.

3. It is an important issue but my impression is that the needing of an enormous mass storage means that science is not done at best. The example of the incredible development of the storage and CPU need of the software normally installed on any PC is instructive: to do the same things done years ago with PC with modest speed and memory one needs GHz clocks and Gigabyte memories!

The most important instruments to deal with such issue are:
First: reduce useless data output by mean of intelligent previous screening.
Second: give the community human resources, more than complicated software.
The scientist does not want, and cannot, handle with everything.

4. The Grid is for now a vague concept. Scientists will take advantage of that when good prototypes will be tested and will show real advantages on the present situation. The most crucial point is communication speed. Every area of modern research will need powerful computing systems, and this is a clear trend in the most evolved countries.

5. The scientific community is not always willing to accept the sharing of knowledge, resources, tools and data, as proposed by the Grid idea. It is in any case often demonstrated that share knowledge, resources etc. is a winning concept in fundamental science.

Issues like knowledge property, data access etc. can be faced by international rules that, anyway, should be improved and adapted to the new needs.

6. Virtual organizations, distributed workshops (e.g. Access Grid), or distance learning can be useful tools for the development of science and, more generally, of knowledge if cleverly organized and thought. The way to create virtual organizations should be determined by a clear preliminary definition of the aims and scopes.

The obvious benefit is: not to do everything by themselves.

Obviously specific infrastructures are required to make them more and more effective. In particular large band-width of data transmission (I include in that efficient video-conferencing, etc.). Easy and reliable (stable) software and hardware.

7. An effort should be made in order to make science and scientific results more comprehensible to people. Actually, technology should be both serve to get advanced insight in nature and to release this knowledge to humans.

The HPC centres must transform from places where one can get (often hardly) just computing (hardware) resources to places where one find "solutions" to problem that must be attacked via computing. So, human resources are needed mainly (support, expertise, scientific collaboration…). Big computers and large mass storage result to be a waste of money without investments to make these facilities really usable.

6.3.5 Interview with Jean-Christophe Desplat, EPCC , University of Edinburgh

Dr. J-C Desplat works as applications scientist at the EPCC Supercomputing centre of Edinburgh. His current interests include management responsibilities for a number of European activities such as the HPC-Europa project (Transnational Access co-ordinator for the HPC-Europa Consortium), the ENACTS network, and EPCC's involvement in TT@MED, the accompanying measure for the EC Medical cluster EUTIST-M. Beside these activities, He is also managing EPCC's involvement in the EPSRC-funded e-Science pilot project the Reality Grid .

6.3.5.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Dr. J-C Desplat can be summarised as follows:

1. Technology will not be driven by a single requirement or application field. However the most important progresses will probably be in the area of data management and information extraction.
2. The market is likely to be relatively mixed. Cost effectiveness is a crucial issue, but also HPC high-end architecture will play an important role for specific applications and users.
3. Data management is a crucial issue. This will be important not just for groups that produce large amount of data but also for those who have lower requirements but can benefit from exchanging, federating and interoperating data.
4. The Grid, first of all, must move away from the development of prototype proof of concept software towards more industrial grade software.
5. Resource sharing is important but some suitable mechanisms/policies need to be defined, and more importantly enforced.
6. Useful tools. Besides obvious advantages (such as cost-effectiveness), they foster a greater spirit of interaction – if not collaboration – enabled by new

technology.

7. HPC centres should spend more effort promoting best practice, and do it more convincingly. The move towards metadata will be a major step forward towards effective collaboration across groups. HPC centres should initiate and foster community led initiatives, bringing their own expertise, to promote best practice in software engineering, code re-use, interoperability, “standardisation” of data format, etc.

6.3.5.2 Interview

1. All of the above (most users simply want “more of the same”). The demand for increased CPU power will grow with Moore’s law as it has always done. There is a bit of a “chicken and egg” problem here: requirements depend on the type of problems you are studying, and most users choose problems which are tractable with the resources they think they can secure. Obviously, there are exceptions.
So a more interesting way to look at this is to think about which of the above will grow at a different rate than it has over the past decade? Or rather, which technologies are likely to make users change their working practices, and which underlying resources will be required by this evolution? In my view, this will be in the area of data management and information extraction (which may allow new working practices like computational steering, interoperability of data across application codes – if not between computation and experiment data – sharing of distributed data sets, information extraction tools, etc.)

Note that different communities are moving in this direction at different rates. It is common knowledge that groups from the particle physics (both experimental and theoretical) have already achieved great progress in this direction, also followed by the astrophysicists and their “virtual observatories”. Other groups in computational chemistry have also started this process, e.g., as part of workshops organised by CECAM in Lyon or seminars at the UK National e-Science Centre and elsewhere. There are a number of barriers we should not overlook though: for instance, computational steering may require more “flexible” scheduling policies at HPC centres (e.g., for co-scheduling / advanced reservation), the current middleware is not yet mature enough for production (and is unlikely to be for another 3 years at least), etc.

2. This question is a bit irrelevant… Looking back at the past 10 years, we have been through a series of trends which almost look cyclic. So what are the main factors? Cost effectiveness has increasingly become an issue, thus constraining vector systems to a niche market. Beowulf clusters (read, DIY Linux clusters, not x86 based systems supplied by mainstream vendors) proved trendy a couple of years ago because of this, but then many realised that 1. The cost of ownership was substantially higher (e.g., for system administration) and 2. Many legacy codes still rely on a replicated data approach, and as such require a shared memory system in order to study moderately complex problems.

The key conclusions I drew from my recent trip to Supercomputing 2004 was that 1. There is a realisation that one size does not fit all; 2. FPGA technology might play an important role in next generation HPC systems; 3. Intel/AMD based systems designed by mainstream HPC vendors will represent the majority of HPC systems in the high-end HPC; 4. “Dedicated” HPC hardware (either custom built like QCD-OC and MD-Grape, or more general purpose like IBM BlueGene/L) are coming back into the market with unrivalled value for money (€/GFlops) and 5. “Commodity” components are increasingly making their way into high-end systems.

To summarise, the market is likely to be relatively mixed (in terms of vendors and technology, which is a good thing for users!), but dominated by Intel/AMD systems (blades or fatter servers). FPGA may play an important step change in terms of (affordable) performance, and a number of vendors have already publicised their intention to allow the integration of such technology within their mainstream products. Dedicated HPC systems like IBM BlueGene/L will also prove very attractive to any group with application codes which can scale on such hardware. Large legacy codes developed around replicated data will still require large memory shared memory servers though.

Another important point linked to this question is how the balance of investment between centralised national HPC centres and organisations (like single universities, or even research groups) will evolve over the next few years.

3. Once again this really depends on which communities you are considering. The particle physics and astronomy have achieved great strides towards an infrastructure which makes it practical to share large amount of data across multiple sites. Progress has been much slower in many other scientific disciplines unfortunately. I am not convinced that the benefits of data management are limited to those groups who are producing vast amounts of data. Other groups with modest storage requirements could also benefit. What is at stake here? Information extraction and interoperability (across application codes) will enable true collaboration across research groups.

So the technical steps are well known. Starting with a community led initiative to define a suitable XML schema representing datasets within their own community, then the deployment of a data management environment like the QCD Grid environment (see www.gridpp.ac.uk/qcdgrid). Other communities may find it preferable to use distributed databases. Recent developments in this area include the OGSA-DAI project (www.ogsadai.org.uk) in which major industrial players such as IBM and Oracle, and academic centres like EPCC are pulling their expertise together to construct middleware to assist with access and integration of data from separate data sources via the Grid.

So how will the community react? Those who can see the benefits of the migration towards metadata / information extraction have already started the process. As for others, they will probably adopt a “wait and see” attitude until the benefits are perceived to be real. In my view, HPC Centres should do more to promote the benefits of such technology within the user community, and I welcome initiatives such as that taken as part of the HPC-Europa data management activity to help educate our users. Indeed, as was pointed out by a number of high-profile personalities from the Grid and HPC community, there is a real need for some “social engineering”. Technological and political barriers are real and should in no way be underestimated, but the sociological aspect should not be overlooked and may represent a greater challenge than the other two (the relatively low level of code re-use in certain communities may only be the tip of the iceberg).

4. To start with, the term “Grid computing” encompasses too many approaches. I will start by excluding the so-called Internet computing (the SETI@HOME type model) and metacomputing, as I do not consider them to have any relevance to the scientific community.

Applications of Grid computing within this community will revolve around two main concepts:
Data centric Grid computing: how to manage/share large amounts of distributed data, where computing is mostly required for data processing / information extraction;
Compute centric Grid computing: here the idea is that you bring added value by clustering a set of (distributed, preferably heterogeneous) HPC resources: such a Grid would provide more choices to users both in terms of hardware and scheduling policies (thus with the ability to cater for both capability and capacity computing within the same “virtual” infrastructure). When implemented, the concept of Grid economics will also be very interesting to allow a better use of these resources (prioritisation): more expensive but with a guaranteed turn-around time for time-critical application, much cheaper for low-priority jobs which may then be used for backfill. In the EU, the EGEE and DEISA projects represent good examples of these two approaches.

A couple of important points here: first, the middleware must be considered as part of the infrastructure, and treated with the same attention as hardware and networking. Second (and also linked to the first point), we must move away from the development of prototype proof of concept software towards a more industrial grade software. There has been a number of interesting talks on this topic, including one by Paul Messina in which he estimated the effort required to develop such robust software as a ten-fold increase compared to the prototype proof of concept. With my user hat on, I would say that the advocates of Grid computing have been promising us (scientific users) too much too early. This had the result of making many users distrustful – if not cynical – about Grid computing. I would be very surprised if the level of enthusiasm for Grid computing (as measured in the first ENACTS report) could be replicated in 2004/2005.

5. I have already discussed this point in an earlier question. One aspect I have not touched though is that of resource providers. Integrating your systems within a Grid includes a number of risks. Besides the obvious “breach of security” (hackers, etc.) you can also add the fact that the lack of robustness of the current middleware may cause substantial disruption to the regular service… which clearly is not a risk anyone with a stringent SLA would be ready to take. There is also a problem a liability. What if an “external” user (but with all the right credentials/authorisations) disrupt your service, either by accident or deliberately. Who is liable? Service operators do not get the same level of protection as they have for their own local users. Some suitable mechanisms/policies need to be defined, and more importantly enforced. To come back to your question, I would refer you back to my comments on the need for some “social engineering” in previous questions.

6. Yes. Besides obvious advantages (such as cost-effectiveness), they foster a greater spirit of interaction – if not collaboration – enabled by new technology. A few groups have already embraced the routine use of AccessGrid technology for regular

meetings. As you suggested however, AccessGrid technology is still missing a number of useful features, in particular with respect to the sharing of applications during a session. However, the AccessGrid developers' community and specific projects (such as the HPC-Europa N2 activity) are making good progress addressing such shortcomings. An interesting development in that respect is the emergence of cheap personal AccessGrid nodes – known as PIG (Personal Interface to the Grid). So you can now build your own personal desktop AccessGrid node for under €5k.

7. At the risk of sounding patronising, I would say further "education". The sociological barrier is clearly underestimated and might prove more time-consuming to address than its political and technological counterparts. HPC centres should spend more effort promoting best practice, and do it more convincingly (i.e., need for greater engagement). The move towards metadata – if carefully thought out – will be a major step forward towards effective collaboration across groups. HPC centres should initiate and foster community led initiatives – bringing their own expertise – to promote best practice in software engineering, code re-use, interoperability, "standardisation" of data format, etc. Organisations such as the UK e-Science Institute are already playing such a role. Personally, I would like to see HPC centres increase their participation in workshops and seminars organised by European organisations like CECAM. But obviously, this is where the issue of funding comes in…

6.3.6 Interview with Jari Järvinen, CSC - Finnish IT center for science

CSC, the Finnish IT Center for Science, was represented by Dr. Jari Järvinen at the interview. Jari Järvinen works as the Scientific Director at CSC. CSC, located in Espoo, Finland, provides researchers with Finland's widest selection of scientific software and databases and Finland's most powerful supercomputing environment. CSC provides services for the Finnish academic research community as a whole, plus research institutions and industrial companies.

6.3.6.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Dr. Jari Järvinen can be summarised as follows:

1. More and more CPU power and storage capacity will be needed in future. Also, tools for data analysis, data mining and visualisation are important.
2. There are two "winning" architectures depending on applications the scientists use: Traditional supercomputers, e.g. shared memory systems, and low-cost clusters (Beowulf clusters) with efficient processors with good interconnections
3. How to analyse the data is one of the most important issues. The scientists in various fields need to co-operate with computer scientists who are experts on data analysis.
4. Scientifically and technologically Grids are very possible in future. But there has to be agreements on political level between nations in Europe on usage policies etc. In Finland, physicists are among the first real Grid users. Other fields in Finland, not so active this moment, are astronomy, earth science and biosciences.
5. The scientific community will not have any problems to share knowledge, tools and data; neither knowledge property or data access will be problems. But what comes to nationally funded HPC resources, politicians and decision makers may

not be ready to share resources at this moment.

6. Access Grid type technologies can be used for organising meetings and workshops, and they can also be useful for educational and training purposes. Access Grid type tools help communication to certain extent, but they cannot replace personal communication. To make these tools more effective, video screen and a talking head are not enough; you should combine lectures, books, a talking head, databases and literature together.
7. There will always be new and better supercomputers and faster networks, up to 40 GByte after a few years. HPC centres will definitely exist in the future, too. HPC centres will maintain large supercomputing systems and offer expert services to help scientists to make science.

6.3.6.2 Interview

Jari Järvinen, Scientific Director at CSC, gave the following answers to the ENACTS in-depth interview questions:

1. First of all, there are a lot of requirements in HPC in Finland and Europe and it is very clear there is a growing demand for HPC capacity. E.g. in Finland, the CPU capacity will grow during the next years, but also requirements of our users will double every 14 months. So there are certainly very many fields in computational science which need more and more capacity. Some of the emerging files are for instance biocomputing, traditional fields as physics and chemistry, and also computational medicine is creasing its roll. The CPU power is absolutely the key question here and more and more CPU power will be needed. There will be a data explosion in several fields of science; so we need more storage capacity already in the near future. Also, we need tools for analysing the data. Therefore, I see that data management tools, e.g. data mining and visualisation are important. I see two parallel developments, CPU power requirements and data storage and analysis. For instance at CSC, we should go in direction a data center.

2. There are a lot of “winning” HPC architectures. Traditional supercomputers, e.g. shared memory systems like the IBM we have at CSC, is a winning HPC architecture since the scientists need peak performance and there are a lot of applications which can be parallelised. There are different needs within scientific fields. Depending on applications the scientists have (e.g. in material science), if they need high throughput capacity, they can use very cheap clusters (Beowulf clusters) with efficient processors connected with Gigabit Ethernet. This interconnection is not necessarily as fast as in traditional supercomputers. In e.g. computational engineering, the applications are strongly coupled. They need fast communication between processors and they cannot rely on ordinary clusters. There is no single “winning” HPC architecture, but there are two trends, supercomputers and low-cost clusters, depending on applications.

3. First of all, the scientific community will produce a huge amount of data from experiments and also from computations. It is the role of supercomputing centres to store this huge amount of data.

The scientific fields will need computer scientists who are experts on data analysis. I think that how to analyse the data is one of the most important issues and it is also cooperation between different types of science. E.g. among bioscientists there are not necessarily so many computer scientists who can analyse the data. So, there are very much requirements to put together different kinds of expertise.

4. The present status of the Grid is very much academic orientated. Like World Wide Web created in Cern, the Grid more or less started from the academia, and it is not a very productive tool at the moment. It is under development, and it has to go through the development phase and still further, through the proof of concept phase. There are a lot of potential Grid projects in Finland, Europe, USA and Japan. They are producing step-by-step an environment which can be used first in science, we talk about e-science, then in e-health for e-governments, and finally there will be e-business. But it will take a lot of time to get there.

In general I think that Grid is an emerging field. For instance, if we think about funding in Europe and in the US, there is a huge amount of funding available. This is a sign that also decision makers believe in Grid technology. Grids are also one part of the ERA, European Research Area. E.g. in the 6th Framework Programme there are a lot of investments in the Grid. I believe in Grid technology, it is coming, and it is widely accepted in the scientific community.

In Finland, we have a pilot Grid project related to material science; I think that physicists are among the first real users. Other fields in Finland, not so active this moment, are astronomy, earth science and bioscience.

About threats of Grid technology, I think that scientifically and technologically Grids are very possible in the near future. But if we think about Computing Grids, the main issue is how to make political agreements between nations in Europe. An example: CSC is a nationally funded organisation offering HPC resources. What is the policy how a researcher can use CSC’s capacity? And how can a Finnish researcher use resources abroad? There has to be agreements on political level.

5. The scientific community shares knowledge when they are producing scientific articles for instance. Therefore I think about this very positively, I think that there will not be any problems in the scientific community to share knowledge. I definitely believe that the scientific community is ready to share tools and data. There are a lot of examples, for example databases which are available world wide already. I think that knowledge property and data access will not be problems in the scientific community.

I think that the most important resources are nationally funded, typically in HPC centres. I am not so sure if politicians or decision makers are at this very moment ready to share resources.

6. Science is international and collaboration is a very essential characteristic of it. Research groups worldwide are more or less virtual organisations. Of course Access Grid type tools help communication to certain extent, but you cannot replace personal communication with Access Grid technology. I mean that the human interface is always very important. Of course you can organise meetings by using Access Grid technology, and it can also be useful for educational and training purposes. Also, you can partially organise workshops by using Access Grid technology. At CSC we have had some lections from abroad. For instance, we had a speaker from UK and a videoconferencing connection to UK and we showed the lecture on the screen.

Distance learning is a very acute theme in Finland also this moment. You may study basic courses at any time and at any position, and it that sense it helps you a lot. In Finland this is mainly related to basic courses and to Virtual University, but still I'd like to emphasize that distance learning is a tool helping to study and make research but nothing else. These are possible and very promising in a sense, but I'd like to emphasise here the role of human interface – these technologies cannot entirely replace the human interface.

Video screen is not enough for creating a workshop environment. To create an environment you should combine lectures, books, talking head, databases and literature together. If this kind of an environment was created, I see it would be more beneficial than just a talking head.

7. Concerning the technology, we are in a very fast moving train, so there will always be new and better supercomputers. New installations are definitely ahead, after 4-5 years there will be again a new supercomputer. Also the networking aspect is very important. Now we are talking about the European standard of 10 GByte. It will remain on that level for some years but after that it will increase to 40 GByte or so. In all fields there will be developments.

As a single national HPC centre, CSC has been very successful. For example, if you think about high performance computing and peak performance they have been higher than in countries with many HPC centres. I think that HPC centres will definitely exist in the future, too, and they will maintain large supercomputing systems. I hope that there will be more scientific expertise in this kind of centres, I'd like to emphasise the role of human resources, to help scientists to make science.

6.3.7 Interview with Kari Laasonen, University of Oulu

Professor Kari Laasonen represented the University of Oulu, Physical Chemistry, at the interview. His main research activities are in the fields of computational chemistry and synthetic coordination chemistry. Located in Northern Finland, just 200 kilometres south of the Arctic Circle, the University of Oulu focuses in biotechnology, information technology, northern issues and environment.

6.3.7.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Professor Kari Laasonen can be summarised as follows:

1. The scientific community will have an endless need for more CPU power. When the data sets get bigger, the need for data analysis and visualisation tools will increase.

2. We will see several types of HPC architectures, each of them most suitable for certain types of calculations. Low cost clusters are suitable for certain applications, but if there is a real need for parallel computing, parallel machines with good connections between processors are needed.

3. Many research groups will use huge data sets. Easy and reliable access to them is important. In a Data Grid type solution the data will be stored in distributed

storages in several geographical locations. Grid type solutions to handle big data sets are being developed.

4. The Grid can balance the computing resources, but it cannot replace the traditional supercomputers. In Finland, we have an interesting pilot project, the Material Sciences Grid, now being installed, but its capacity will be moderate compared to supercomputers. For data intensive fields, Grid type tools to share the data may bring something completely new. In Finland, these data intensive fields could be remote sensing, forest industry, etc.

5. Traditionally, the scientists have been used to share their knowledge and data. The bureaucracy may get harder by taking the Grid in use, because there has to be agreements how to use the data. Everyone wants to take their own share, and national resources will be defended.

6. Virtual tools can make it easier to work in national and international networks. To speed up the adaptation of virtual tools, both technological and human resources are needed. When organising meetings, possibility for virtual participation should be offered for participants in remote locations. These tools should be easy to use at your desktop.

7. The most important requirement is to get more computer capacity to the users, efficient parallel machines especially. It is a natural role of HPC centres to offer computing and storage capacity. The new Grid ideas are good and the HPC centres should take part in building the Grid systems.

6.3.7.2 Interview

Kari Laasonen, Professor in Physical Chemistry at the University of Oulu, gave the following answers to the ENACTS in-depth interview questions:

1. The future HPC requirements of the scientific community will be mostly CPU power and to some extent analysis and visualization tools. The CPU power demand of the computational science is endless. With more computer capacity, bigger, more realistic and more important problems can be studied. When the data sets get bigger, the need for data analysis and visualisation tools will increase. But the need for more CPU power comes first; tools are secondary.

2. We will see also in future several types of HPC architectures, each of them most suitable for certain types of calculations. To certain applications the low cost clusters are very good. But if there is a real need for parallel computing, parallel machines with good connections between the processors are needed. Clusters have come to stay, and more clusters will come. But researchers from various fields will have differing needs, and we will also need real, effective parallel machines.

3. Huge data sets are often generated by big organisations, e.g. telescopes, global satellite systems etc; they are expensive measurements the amount of data is very big. Huge data sets will be used by many research groups, and easy and reliable access to them is important. I see the Data Grid type approach as the most useful. The analysis is usually focused to only part of the data and thus, it is easy to parallelise. The data will be stored in distributed storages in several geographical locations that will have

certain protocols and programs for analysing. Ordinary computers will be capable for the most of the data analysis. Technical solutions exist for handling big data sets, but Grid type solutions for big data sets are being developed.

4. The Grid development is in a very early state, and the benefits for traditional computational sciences are not that great. I think that the Grid cannot make a break-through in computational science, at least in the near future. The Grid can balance the computing resources, but it cannot replace the traditional supercomputers.

In Finland, the national supercomputing environment provided by CSC is powerful, and it will hardly be beaten by the Grid systems. We have an interesting pilot project, the Material Sciences Grid, now being installed on seven university campuses in partnership with CSC. But its capacity will be moderate compared to the supercomputers. For data intensive fields, Grid type tools to share the data may bring something completely new. I think they will appear in the near future. In Finland, these data intensive fields could be remote sensing, forest industry, etc.

5. Traditionally, the scientists have been used to share their knowledge and data. On the other hand, the competition in science is strong, and the scientists have to get the glory of their important findings. But after they have published their findings, they want to share them.

There is a risk that large data sets can be misused or misinterpreted. So I see some problems with the data and knowledge sharing, which have to be solved before taking the Data Grids in use. There has to be agreements how to use the data, how to cite it and who is responsible for the analysis and interpretations. The bureaucracy may get harder by taking the Grid in use. Everyone wants to take their own share, and national resources will be defended. Let’s take an example from the biosciences: If a group finds a new genome and shares their data globally, it is not acceptable if someone else applies for a patent by taking their findings, and gets big profits.

6. I hope the distance learning and workshops will help to organise virtual meetings since it is rather costly and time consuming to travel abroad just to listen a few talks. Living in Northern Finland, it nearly always takes a day to participate a meeting out off your office. I hope that virtual tools will make it easier to work in national or international networks. I strongly hope that the virtual communication tools will be in wider use in near future. To develop these tools, the universities and the network operators must collaborate. When organising meetings, possibility for virtual participation should be offered for participants in remote locations. These tools should be easy to use at your desktop. To speed up the adaptation of virtual tools, both technological and human resources are needed.

7. The most important requirement and expectation with respect to the near future technological development is to get more computer capacity to the users, efficient parallel machines especially. Grid systems will arrive and be suitable for data intensive needs, e.g. the Cern experiments. But the Grids cannot replace the traditional supercomputers. In a country with a small population, it makes sense to concentrate resources.

In my opinion, the computing centres should remain as computing centres also in future. It is a natural role of HPC centres to offer computing and storage capacity. In Finland, the national HPC centre is responsible for offering computing resources. The new Grid ideas are good and the HPC centres should take part in building the Grid systems. The new thing is that access to Grid resources will be wider than it is now; in addition to national users there will be users from abroad. It is probable that big Grid solutions will bring more bureaucracy. I would like to see a functional Grid solution between a few countries before I believe in global Grid solutions.

6.3.8 Interview with Matthias Mueller, HLRS, University of Stuttgart

Dr. Matthias Mueller is the Head of Technical & Scientific Computing Division at the High Performance Computing Center Stuttgart (HLRS) of the University of Stuttgart. His field of interest is dry granular matter. One of the main topics is to develop parallel algorithms to simulate granular materials.

6.3.8.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Dr. Matthias Mueller can be summarised as follows:

1. The scientific community will require a balanced combination of CPU power, data storage, and data management and visualization tools. CPU power is losing its central importance.

2. Different problems require different architectures. There is not a winning choice.

3. Environments which provide not only computing power but also more complete functionalities of data handling are required.

4. The Grid is a research concept not a reality

5. Data and resource sharing depends on the community.

6. Collaborative distributed work will be very important, but reliable and fault-tolerant tools must be available.

7. Specialized computing centres will develop. Collaboration between them is necessary to cover all the possible requirements of the scientific community.

6.3.8.2 Interview

1. The scientific community will require a balanced combination of all of the above. Depending on the specific community, the balance with the best price/performance will be different. CPU power will be less and less important, therefore benchmarks like linpack will loose significance and a list like the TOP500 list will loose expressiveness to judge the value of a supercomputer installation for the scientific community.

2. With multi-core CPUs shared memory systems will be the dominant systems with respect to the number of systems available. They also will be very popular platforms

for new applications that want to achieve speed-up on moderate number of CPUs using OpenMP. Low cost Beowulf clusters will be very attractive systems up to 256 nodes. Distributed systems (“The Grid”) is a solution for a small fraction of the computational problems and also the software solution for access to a large range of systems, including HPC systems. Parallel systems are solutions for the top-end of systems due to increase reliability and serviceability demands and also the driving force for new developments that will later also be available in the commodity systems.

3. The most important instrument is first an environment that does not only look at the compute capacity and second software that can handle the data handling process throughout the complete life cycle of data.

4. The Grid is currently much more a research topic than a solution that can be used by a scientist. It contains some valid ideas but there is too much hype and the Grid is oversold.

5. There can be no general answer to a question like this. It depends on the community or also whether the real value is in the data or rather in the data analysis. For resources it depends on the fact whether sharing means free usage or only flexible access with billed usage. Without a market place trading of resources will be almost impossible.

6. Distributed workshops and distance learning will become more and more important. One requirement is that the tools will be reliable and fault-tolerant. Usage should be as easy as a phone call.

7. The issue of price/performance will increase for the majority of centres. There will be only a very small number of centres that will try to achieve top performance without looking at the cost. To achieve excellent price/performance, tailored solutions for specific domains are required. Since this favours centres that focus on a specific domain an increased collaboration between different centres is required to achieve a compute environment for the whole computational science community.

6.3.9 Interview with Risto Nieminen, Helsinki University of Technology

Academy Professor Risto Nieminen represented the Helsinki University of Technology, Laboratory of Physics, at the interview. His research interests are computational condensed-matter and materials physics, nanosciences and nanotechnology, complex materials and multiscale modelling of their properties and processing, high-performance computing and networking and numerical and computational methods. The Helsinki University of Technology is located in Otaniemi, Finland, where top-level technological research and business meet and whose architecture honours the spirit of Alvar Aalto.

6.3.9.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Academy Professor Risto Nieminen can be summarised as follows:

1. More CPU power will be needed, because the user community is growing and new areas are becoming compute-intensive. There will be a need for sophisticated software for data-intensive computing. Also, the demand for more storage capacity exists.

2. For the throughput volume computing Linux clusters are the winning solution. The evolution in the software area will be towards Linux clusters, too. However, shared memory systems will remain as marginal solutions for selected sub-communities.

3. The communities have to invest more on storage capacity, various types of disk arrays and archiving systems. The scientists around the world should have access to the growing amount of data stored in distributed storages. However, in the near future the solutions for handling distributed data archives will be discipline-based rather than universal.

4. The Grid will develop in a topical fashion arising from the needs of scientific sub-communities. For a scientific sub-community, the Grid will be a joint strategy to share computing technology, resources and joint scientific interests. In Finland, a national Material Sciences Grid is now being installed to make a joint resource for materials modelling and simulation. Other communities having the joint scientific interests and goals are likely to copy this model.

5. It is likely that scientists are willing to share knowledge, resources, tools and data, which is the prerequisite for having a successful Grid architecture. There will be communities and scientists who don’t agree to share their data or who see serious IPR issues, but they won’t need the Grid.

6. The Access Grid type technologies have not yet developed to the point where it would be easy enough to communicate with others. In order for these tools to be useful, you need them at your desktop. However, these tools can offer a useful complement to regular education and training activities.

7. Clusters, possibly combined with Grid access, will offer a technological solution to the increasing demand of CPU access. HPC centres should help the scientific communities to build these clusters. HPC centres should provide the peak performance to the grand challenge type projects. That is very expensive and only makes sense if you do it centrally.

6.3.9.2 Interview

Risto Nieminen, Academy Professor at the Laboratory of Physics at the Helsinki University of Technology, gave the following answers to the ENACTS in-depth interview questions:

1. I emphasise the need for more CPU power simply because the user community is growing and new areas are becoming compute-intensive. Also the demand for peak performance will keep on increasing in terms of so-called grand challenge type projects. There is a very serious need for more CPU power and throughput volume computing. With peak performance I mean the top performance you can allocate for single jobs and single projects. Peak performance has not grown very dramatically

over the years; we have seen most of the growth in terms of volume computing. There are scientific areas such as climate modelling and other complex phenomena where the demand for peak performance and shorter turnaround times is going to be very important.

More storage capacity will be needed. In high energy physics, environmental modelling, astrophysics, cosmology etc. we have these mega projects now on the way that require very large storage systems. The demand for more storage capacity perhaps is easier to meet, but anyway it is there. Certainly we need sophisticated software for data-intensive computing, for example so-called middleware. Data analysis and visualisation tools are of course important. Visualisation in particular has become such routine that I don't see any important bottlenecks or serious problems, but of course we will see continuous development also in those.

2. In the present situation, there is no doubt that for the throughput volume computing Linux clusters are the winning solution; they offer by far the best price-to-performance ratio for the scientific community. This is in particular true in such areas where computational science has established itself and has strong traditions, in physical sciences and more and more also in biological sciences. The volume of the scientific computing is already now clearly on Linux clusters. We have witnessed during the last few years a very dramatic decrease with moderate costs of the communication switches in clusters, e.g. Gigabit Ethernet cards are cheap, standard and routine. It is very easy now to build cluster-type solutions that have actually quite considerable bandwidth among the nodes.

Shared memory systems will offer the best peak performance for selected applications. We see now some kind of re-emergence of sophisticated shared memory systems designated for the HPC area, the rebirth of Cray for example. Shared memory systems will always be marginal solutions; they will offer very sophisticated service and power for selected sub-communities. But for some strategic needs there will always be countries and organisations that will invest in these very expensive shared memory systems.

I don't think that the researchers from various fields will have differing needs in the end. If you look at the present situation, you can say there are certain types of 3rd party software, which only run on certain machines. The evolution will be in the software area towards Linux clusters.

3. The communities certainly have to invest more on storage capacity, various types of disk arrays, archiving systems etc. This technology is growing in importance. I don't think there is a major technical worry there; these solutions are pretty mature and developed.

The second issue is management, data mining and utilisation of data and databases, which is very much discipline-dependent. In the high energy physics area, the community is coherently investing in producing management and data mining tools to share the data, even it is stored in a distributed fashion. Other communities, for example in bioinformatics, astrophysics, cosmology, materials physics etc, will certainly repeat this. There will not be a universal solution at least in the near future for handling all types of distributed data archives, but the software tools will be more discipline-based and developed within that particular community.

The amount of data will increase quite dramatically. We have these mega experiments coming online, like at Cern and some other experiments, which will produce very large amounts of data. The data will be stored in distributed storages almost globally. The challenge will be to guarantee access to this data in a meaningful way so that scientists around the world can process and analyse this data.

4. Grid technology and associated ideas are worth pursuing and will be certainly pursued, although there is not a clear definition what a Grid actually is and various people mean slightly different things with it. Grid is here to stay. Grid will develop in a topical fashion. Grids will be topical and they will arise from the needs of scientific sub-communities. There will be a High Energy Physics Grid, Materials Grid, Astronomy Grid etc. The Grid solutions are best developed for example in the community around Cern experiments. There they have a relatively well defined the problem that they want to solve using distributed technology, distributed storage and data processing capabilities, because the amount of data and data analysis is very large and it makes no sense to handle everything centrally.

Grid becomes a joint statement, joint strategy for that scientific community, who share computing technology, resources and joint scientific interests. They are ready to commit their data, ideas and effort to a joint purpose; this is a requirement for a useful Grid, at least in the near future. Given that basis of collaboration and common goals, it is natural to link distributed resources together and develop software which enables users to access that resource wherever they are geographically. At best this will be very robust technology, not vulnerable to failures in single nodes for example, and you can utilise the resources the best possible way.

In Finland, the high energy physics community, not very large but active, are participating in the Cern related developments, together with the other Nordic countries. They are already demonstrating the capabilities of the Grid architecture in the upcoming analyses of the Atlas and CMS experiments at Cern. These experiments will produce enormous amounts of data that have to be analysed in a distributed fashion and the Grid will provide that possibility. In Finland we have the Material Sciences Grid, which is now being installed. Nearly identical Linux clusters on 7 university campuses will be put together to make the joint resource for materials modelling and simulation, sharing common application programs, sharing data and tools etc. I am absolutely sure that other communities, having the joint scientific interest and goals as a starting point, will copy this model.

5. The scientific community has to have the willingness to share knowledge, resources, tools and data; this is the prerequisite for having a successful Grid architecture. If you don’t like that idea then there is no reason why you should go for a Grid type solution. It does not mean you have to share everything or collaborate in every aspect of your work, but to certain extent you have to be able to open your books and drawers, tell people what you are doing and also share your results with them. I think there are many communities and sub-communities willing to do it, because it is the only way for them to reach the major, strategic goals. There will be competition within the Grid community but the starting point is the willingness to share knowledge and data.

The questions about IPR and other copyright issues are secondary, because if you feel there is a serious IPR issue you will not be interested in joining the Grid. One can foresee these issues may appear later after the Grid has started and people have started sharing data, but I think this is a fairly hypothetical question. You enter the Grid collaboration voluntarily and the community has to agree on these issues before anything happens. It is likely that scientists are willing to share knowledge and data, at least in high energy physics. We have many examples already, not only Grid solutions in place, but in the scientific community also the open-source thinking is common. We develop software in the open-source area routinely and it is the winning strategy most of the people are ready to subscribe to. But there will be communities and scientists who don't agree to share their data, but they don't need the Grid. Those communities who join the Grid will win from that kind of collaboration.

6. From the research point of view, Access Grid type solutions or distributed workshops have not yet developed to the point where it would be easy or routine enough to communicate with others. In most cases you still have to go to a special area or room and set up e.g. Access Grid to organise a meeting with your colleagues at other geographical locations. I have taken part in that kind of meetings every now and then, but they are not routine, and they are relatively minor activity.

In order for Access Grid and teleconferencing tools to be useful, you need them at your desktop. You come to your machine and can start talking to your colleagues wherever they are; at that point they are very useful, like email or other routine applications of the Internet are today.

These kinds of tools are, however, important for learning and teaching activities. Many countries, including Finland, have put a lot of effort to develop tools for distance learning and virtual universities. Progress has been slower than hoped; the acceptance of these tools seems to take a lot of time and developing content for virtual organisations is a very heavy process. For special solutions like voluntary-based adult education, distance learning in special circumstances can be very useful and hopefully further developed. But they are marginal and they won't offer an alternative to e.g. normal university-level education. However, they can offer a useful complement to the tools used for regular education.

7. The most important requirement is access to CPU power. There is a serious shortage in Finland of the CPU power now. In Finland we have a good history in computational research and during the 1990s we were able to lay a very solid groundwork and demonstrate success in this area. This has brought with itself an expansion in the scientific activities, the computing has grown into totally new areas, the user communities are much larger and therefore the demand for CPU power and related services has increased very dramatically.

As a single national HPC centre, CSC has been active in triggering the growth in demand in the past and it has been very successful in raising the standards and developing tools and awareness of the scientific community. I feel that the technological solution to the increasing demand of CPU access will be in the cluster technology, clusters possibly combined with Grid access. The role of the HPC centre will certainly change; I think we should go towards more distributed solutions in the future. We should help the scientific community to build these clusters at campuses, link them together into Grid type solutions, form virtual communities, and discipline dependent communities. We are now seeing in the Material Sciences Grid that this is good development, and the HPC centres should do everything they can to help this solution.

We still need, at least in Finland, a lean and mean HPC centre to provide the peak performance, or peak capacity, which will not correspond to a very large fraction of the total volume of computing but it will be very specific to the most sophisticated users and grand challenge type projects. I think the HPC centre should focus on high performance, really top-end computing. That is very expensive and it only makes sense if you do it centrally.

6.3.10 Interview with Kai Nordlund, University of Helsinki

Professor Kai Nordlund represented the University of Helsinki, Computational Physics unit at the interview. He works in the Accelerator Laboratory at the University of Helsinki, Finland, his current research focus being in the fields of materials science, nanotechnology, laser physics and computational materials science.

6.3.10.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Professor Kai Nordlund can be summarised as follows:

1. There will always be a need for more CPU power. There are many fields that generate lots of data; this moment there is not enough storage capacity available. The need for data management tools will increase. Visualization tools can be extremely useful but their relative importance will not change in future.

2. The future "winning" HPC architecture will be distributed memory architectures. The trend is large clusters with common, relatively cheap processors and efficient networks between processors. Shared memory systems and vector machines are too expensive.

3. Price-to-performance technical solutions exist to store, manage, mine and analyse the produced data. The biggest challenge nowadays is to develop solutions for very big data sets.

4. Grids will be used to get distributed computer capacity and data within the scientific community. For example, in the NorduGrid we can already see that the Grid is a useful tool. To move beyond the scientific community, the Grids should be reliable and easy enough to use.

5. I think that the scientific community is definitely willing to accept the sharing of knowledge, resources, tools and data, as proposed by the Grid idea. However, many commercial companies will never share their work.

6. Videoconferencing and Access Grid could facilitate the daily work and routine work of the multinational research groups. Videoconferencing can never be as

nice as meeting people; the research groups are likely to make new contacts and the most important scientific developments in person.

7. Cluster computing is becoming more popular. The relative importance of data management and data storage solutions may increase. HPC centres have an important role in maintaining the knowledge and educating the local groups on the Grids. Data storage and data mining are natural roles for HPC centres.

6.3.10.2 Interview

Kai Nordlund, Professor in Computational Physics at the University of Helsinki, gave the following answers to the ENACTS in-depth interview questions:

1. CPU power is the clear need; there will always be a need for more CPU power. The need for storage capacity is increasing. There are many fields now, e.g., physics, bioinformatics and atmospheric sciences that generate lots of experimental simulation data that needs to be stored and processed. This moment, there is not enough storage capacity available. Data management tools may be important for research fields that need data gathering, currently not necessarily high energy physics, but the need will increase. The need for visualization tools will always be and they can be extremely useful but most probably their relative importance will not change in future. Still, CPU power will be the most important thing.

2. Researchers from various fields may have differing needs, but the future “winning” HPC architecture will most probably be distributed memory architectures. They will be based on relatively cheap components and efficient communication. The trend is clear: nowadays you can see large clusters with common, relatively cheap processors with efficient networks between processors. You can also see that relatively cheap decent networks (like Myrinet) have become better and better. Shared memory systems and vector machines are nice but too expensive, and there will be very few scientists who need them so much that they are prepared to invest the extra cost needed.

3. Price-to-performance technical solutions exist to store, manage, mine and analyse the produced data. The biggest challenge nowadays is to develop solutions for very big data sets. Big projects, e.g. the Cern DataGrid, are under development. They know what their needs will be, and they are working hard to find solutions. It is not 100% certain yet whether the concept with several data layers will work, but there is a big need for solutions of this kind. Much of the technical development in high performance computing will have to be directed to finding solutions in this field.

4. Grids can definitely be useful for the scientific community. Grids can be used to get the distributed and redistributed computer capacity, for example, in the NorduGrid we can already see that the Grid is a useful tool. Grids may remain as limited solutions used by traditional high performance computing and people needing manageable huge data sets distributed over several places.

Commercial companies are currently promoting the same concepts, not necessarily calling them Grids but e.g., on-demand computing. On-demand computing is a special concept in the Computing Grids, and to some extent they already exist in the traditional mainframe computing environment. Whether the Grids will become something way beyond the scientific community, a solution used for almost all computing, is a big question. As soon we move beyond the scientific community the challenges are also bigger. The scientific community needs a lot of computer capacity and they are prepared to work a little bit extra to get the Grids work. Lots of development is needed to make the Grids available to almost everyone in the society. For that purpose the Grids should be reliable and easy enough to use. In 5 years we will be wiser.

There are ideas about Access Grids, enabling further collaboration and ideas that the computing infrastructure will lead to new concepts of collaboration. I'm not sure if these can promote further communication between scientists who are already good at changing information and data at every level.

5. I think that the scientific community is definitely willing to accept the sharing of knowledge, resources, tools and data, as proposed by the Grid idea. Most researchers are extremely open, and after the results are published they want to share them. Of course they want to get credit on their own idea. However, I would say there are almost no knowledge property issues that would be problematic. However, there are a lot of commercial companies that will never share their work.

6. What comes to virtual organisations, distributed workshops (e.g. Access Grid), or distance learning, I'd like someone to show me the killer application. Maybe that just needs some imagination to know what that could be. It is difficult to see how new solutions could bring something fundamentally new. Technical solutions to enable videoconferencing could be quite useful, but at the same time there is a completely different development competing with videoconferencing: cheaper flights. Videoconferencing can never be as nice as meeting people. Senior people, who are not so used to videoconferencing or to use computers, prefer to meet each other. Also young people in science, who are as used to computers as anyone can be, do the same. Although you use videoconferencing, you want to meet people a couple of times a year.

At the Accelerator Laboratory of the University of Helsinki multinational research groups are very common, half of the scientific publications are made in cooperation with someone from abroad. Videoconferencing and Access Grid could facilitate the daily work and routine work of the multinational research groups. I'm pretty sure that the groups are likely to make new contacts and the most important scientific developments in person.

7. One expectation the scientific community concerns the network computing. Most scientists, even computational scientists, have not taken any definitive views with respect to the near future technological development. They rather wait and see. Data storage issues have become more important to a part of scientists. Cluster computing is becoming more popular. The relative importance of data management and data storage solutions may increase.

HPC centres should and can respond to major changes in the overall picture. They can have specialists who know the technical issues of Grids, and the HPC centres have an important role in maintaining the knowledge, because universities cannot guarantee continuity, and they can educate the local groups. Data storage and data mining are natural roles for HPC centres that gather the knowledge to maintain the hardware and have the specialized, expensive hardware.

6.3.11 Interview with Sven Stafström, Linköpings universitet

Professor Sven Stafström represented Linköpings universitet, Department of Physics and Measurement Technology, Computational Physics group, at the interview. His research interests are computational condensed-matter and materials physics and nanosciences. In particular he works on a wide range of physical properties of large carbon based system such as conjugated polymers, fullerenes and nanotubes. Professor Stafström is also the director of the National Supercomputer Centre (NSC) located at Linköpings universitet. NSC is one out of three national supercomputer centres for researchers in Sweden. NSC provides leading edge high performance computing resources and support to academic users throughout Sweden and NSC's partners SAAB and the Swedish Meteorological and Hydrological Institute (SMHI).

6.3.11.1 Summary of Key Ideas

The most important ideas conveyed at the interview with Professor Sven Stafström can be summarised as follows:

1. All HPC users will require more CPU power. Storage, data management and visualization tools are important for most users.

2. The winning HPC architecture will most probably be commodity clusters with high speed interconnect. Users will have different needs depending on how tightly coupled the applications are; some applications require shared memory resources.

3. Organizational issues and international collaboration will be very important when storing and analysing huge data sets.

4. The Grid technology has the potential to simplify the work and to make the usage of HPC resources more efficient. The Grid environment is likely to become the dominating user environment in the future. In Sweden, the SweGrid project has already reached the production status for many research groups.

5. Sharing of knowledge is a central theme in the Grid idea. In the future there will be both open and commercial software tools available. Databases are to a large extent open, but commercial interests may restrict the access to data.

6. The next development in information technology might be towards distributed workshops with good visual and spoken contact between different sites. This will require higher speed for data transfer and more advanced visualization solutions. These solutions will offer a faster, cheaper and more environmentally friendly way to communicate.

7. Near future technological development in terms of faster interconnections etc. has to be matched by changes in software. HPC centres can play an important

role in giving users more advanced application support, e.g. in code optimisation, Grid usage etc.

6.3.11.2 Interview

Sven Stafström, professor in Computational Physics at Linköpings universitet and the director of the National Supercomputer Centre (NSC), gave the following answers to the ENACTS in-depth interview questions:

1. All HPC users will require more CPU power. Storage and data management tools will be critical for a few groups, e.g. in the fields of bioinformatics, high energy physics, medical visualization, and climate modelling. Visualization tools are important for most users but for the majority of these users this is taken care of at home, not at HPC centres.

2. Looking at the trend at the top 500 list, clusters have now more than 57% share, three years back the share was less than 10%. The winning HPC architecture will most probably be commodity clusters with high speed interconnect. Propriety solutions will be available at a few centres with specialized users. Users will have different needs depending on how tightly coupled the applications are. Some applications, e.g. Monte Carlo, can run very efficiently on commodity interconnects whereas e.g. CFD problems require high speed interconnects and even shared memory resources.

3. To store, manage, mine and analyse huge data sets, organizational issues and international collaboration will be very important. A hierarchy of storage facilities have to be set up with special software to take care of the data management. In many cases the data will be stored at a specific site and the codes for analysis will be transferred to run on these sites.

4. The Grid technology has the potential to simplify the work and to make the usage of HPC resources more efficient. The Grid environment is likely to become the dominating user environment in the future.

In Sweden, the SweGrid project is operative and in full use. SweGrid is a test project but it has already reached the production status for many research groups. Naturally it will take some time for all users to reach this stage. Advanced scripts, for instance, are hard to implement on the Grid.

5. Sharing of knowledge is a central theme in all science and we don’t think that the Grid idea will change this situation. As an example of sharing the resources we can mention that Sweden and Norway are discussing this issue. Probably one has to start from smaller user groups and then extent to wider communities. There will be both open and commercial software tools available. We will probably live in a situation where somewhat immature codes are available more or less free of charge, whereas some companies invest in code development and seek revenue in the form of commercial software. Databases are to a large extent open but in cases where there are commercial users, there will most probably be commercial interests that will restrict the access to data.

6. Information technology has during the past decade completely changed the way scientists, as well as other users, search information. All scientific journals are now on-line and search engines can be used to find the requested material.

The next development in information technology might be related to direct contacts between scientists in the form of distributed workshops with good visual and spoken contact between different sites. This will require higher speed for data transfer and more advanced and more inexpensive visualization solutions than what we have today, but there is no doubt that the development in technology will bring us there within a decade or so from now. The major benefit will be that the exchange of information can occur much faster and more informal. In the long run these solutions will be less expensive and they will offer a more environmentally friendly way to communicate.

7. There is still a large gap between the maximal performance that the technology can provide and the performance of individual codes. Near future technological development in terms of faster interconnections etc. has to be matched by changes in software.

HPC centres can play an important role in giving users more advanced application support, e.g. in code optimisation, Grid usage etc. Larger research groups (centres excellence) that use HPC should be linked to centres and in such way establish personal contacts between users and HPC experts.

7 Conclusions and Final Remarks

In this report we have presented the results of a research which aims at analysing the requirements and desiderata of European HPC users with respect to high-end computing resources, applications and data management tools. Moreover the survey outlines how users perceive emerging technologies and how these can affect their research and development work.

The research is based on a questionnaire dedicated to users of HPC facilities and datastores and Grid users, in particular researchers and scientific software developers who require medium/large computing resources, so that they can provide a meaningful feedback on high-end CPU devices and infrastructures.

Eleven in-depth interviews to selected representatives of computational science research groups or organisations in Europe have been collected and analysed in order to provide a wider perspective vision of how technological resources should evolve in order to fulfil the expectation of the research community.

7.1 Conclusions on the User Questionnaire

The questionnaire was answered by 125 users from eighteen European countries, mostly university researchers, representatives of a much larger scientific community. Most of the participants are part of medium (2-10 members) or large (11-50 members) research groups. It must be noticed that the majority of these collaborations are local; only a third of the research groups account for international collaborations. This is mainly due to the difficulty of having everyday remote working sessions. The collaborative environments technologies can be crucial to improve the diffusion of transnational working groups and the spread of skills and knowledge across Europe.

The average present-time HPC user which emerges from the Questionnaire Part 2 shows that there is still a “traditional” approach to computer sciences. Most of the participants use local resources (workstations, departmental servers) with small (1 to 4 processors) or medium (8 to 32 processors) configurations. The main concerns are related to the speed of a single CPU and the memory size, rather than having plenty of distributed resources. The access to the computing platforms is mostly via ssh connection, rather than more sophisticated methods like web portals. This can be due mostly to security concerns. The most common operating systems are Linux and various proprietary Unix flavours – Aix, Irix etc. Windows is getting a growing success, even though it is still quite little diffused in the research community. Other products like MacOS are quite uncommon. A large fraction of the researchers’ computing related work is dedicated to code development. Commercial or freeware codes are not common in the scientific community. This is due both to the high specificity of many problems, which require dedicated algorithms and codes, and to certain scepticism towards commercial software. Self-made or self-modified codes, starting from previously home-made programs, are the common choice for most of the researchers. Very little space is given to commercial or freeware applications. Specialised scientific libraries are frequently used, since they are highly optimised, precise and accurate tools to perform standard tasks (like array and linear algebra operations, FFT etc.). A traditional approach to numerics is once more confirmed by the choice of Fortran and Fortran 90 as programming languages. However, also other high-performance languages, in particular C and C++, are starting to diffuse, emphasising the openness toward the experimentation of different and new technological opportunities. This is confirmed also by the interest toward open source products, which usually do not represent completely stable and easy to use tools, but can provide the basic components to develop new applications.

The data seem not to be a major concern for researchers. Most of the applications are computing intensive, but the amount of data that are produced is rapidly growing. However their storage, management and even analysis are considered as a secondary problem. Results are usually stored in files, with no particular organisation and often with no standard format (like CGNS, FITS, HDF). Usually either raw binary files or ASCII tables are used to save data. This can represent a strong limitation for data exchange even inside the same research group. But, first of all, it can be a critical challenge for the standardisation and interoperability effort of the international community and, more generally, for diffusion of the knowledge, cooperation and best exploitation of resources.

The crucial role of the computing power emerges also from the analysis of the users' interest and involvement in Grid related issues. In fact, the main attractive is to have a large aggregate computing power available, with much less concern of its architecture. This is confirmed also from the other points the users have indicated as the most critical in a distributed system, that are high network bandwidth, principally to download data, and efficient schedulers, to maximise the throughput of the workflow. Other opportunities, like portals, shared file systems on distributed platforms, large storage capacity, data management via databases, collaborative working session, are perceived as much less interesting and attractive. The Grid, since it is still in a development phase, is seen as an unfriendly environment, mainly due to lack of stable API, incomplete documentation, difficulties in the management and in the development of suitable applications. Nevertheless, a large fraction of the users is willing to contribute to the Grid infrastructure development, if proper help and support is provided. It is also interesting to notice that more than 27% of the users have already been involved in Grid related research projects. Finally, it is very encouraging that a large majority of the users is willing to share their codes and, under proper conditions, their data and results. However, those who do not want to share resources justify their choice with the fact that these are too specific of their work and therefore useless for the community.

7.2 Conclusions on the In-Depth Interviews

The in-depth interviews have outlined the following results.

The need for more CPU power was emphasised by most representatives. Some of them considered the need for more storage capacity as important as CPU power, and some saw that data storage and data management have become more important than CPU power. The importance of data management tools is likely to increase and the development of data analysis, data mining and visualisation tools will continue. The future "winning" HPC architecture for the scientific community will be clusters for the throughput volume computing. Clusters, possibly combined with Grid access, will comprise common, relatively cheap processors and efficient networks between processors. Also the evolution in the software area will be towards clusters distancing from platform-specific applications. Shared memory systems, efficient parallel machines especially, will remain as solutions for selected sub-communities who need peak performance. In some answers the price of the shared memory systems and vector machines was considered too high for the future systems. The amount of data will increase in the future. Many research groups will use huge data sets stored in distributed storages. Easy and reliable access to the data sets is important and it requires international collaboration**.** Technological solutions have to be developed to handle very big data sets.

The Grid is here to stay and developing according to all the interviewed, but opinions of its development vary. Grids can be useful for the scientific community, or Grids may remain as limited solutions used by traditional high performance computing and people needing manageable huge data sets distributed over several places. For a successful future of Grid infrastructures, a lot of development is needed to make the Grids more reliable and easy enough to use. There are interesting pilot projects going on all around EU. The interviewed agree that the scientific community is willing to accept the sharing of knowledge, resources, tools and data, because it is the prerequisite for having a successful Grid architecture. However, there will be both open and commercial software tools available. Databases will be open to a large extent, but commercial interests may restrict the access to data. The real challenge is collaboration on both the political and technical level. More agreements will be needed how to use the data and resources.

Virtual organisations, distributed workshops (e.g. Access Grid), or distance learning are already useful tools in educational and training activities. Today, the technology is still difficult and requires usually a set of non-standard equipment and software and special skills to operate. In order for these tools to be more useful for the scientific community, both technological and human resources are needed, e.g. to have easy access from your desktop. These tools can enable faster, cheaper and more environmentally friendly communication.

The developments of cluster computing and data storage solutions are seen as the most important expectations of the scientific community with respect to the near future technological development. The HPC centres will have an important role in helping the scientific communities build cluster solutions and giving users advanced application support, e.g. in code optimisation and Grid usage. HPC centres can become the backbone of a true Grid infrastructure offering the scientific community an easy and sound access to HPC and datastores.

7.3 Closing Remarks

This Survey of Users' Needs is the last survey within the ENACTS project. Comparing the results from this study with the first ENACTS study, Grid Service Requirements in 2001, the following notices have been taken.

7.3.1 What do users expect from a Grid?

More computing power and possibilities to study larger scale problems

Computing power still has a crucial role in the involvement in Grid related issues. More CPU power will be needed, because the user community is growing and new areas of research are becoming computing intensive. The deployment of small local Grids has increased considerably. This far, HPC centres have purchased bigger, more efficient machines that have enabled the study of larger scale problems more than current Grid systems in Europe have. Distributed Grid solutions with good interconnections may offer a cost-effective way to tackle larger scale problems in future.

Requirements differ across different groups or research areas. Grids will arise from the needs of scientific sub-communities and the direction is rather towards topical Grids than towards one global Grid.

Increase opportunities to share codes and applications

The Grid is in development stage and this moment the Grid has not really facilitated the sharing of codes or applications yet. Users are, however, willing to contribute to the Grid infrastructure development, if proper help and support is provided.

Grid user support, documentation and application interfaces should be developed. Virtual tools for education and training can enable faster, cheaper and more environmentally friendly communication. To enhance the adaptation of virtual tools, these tools should be more user-friendly.

Ability to share distributed data

There have been Grid experiments that have shown that the Grid, particularly Data Grids, can facilitate data analysis and visualisation.

The growing amount of data that are produced puts challenges. Standardisation and interoperability efforts of the international community are needed for diffusion of the knowledge, cooperation and best exploitation of resources.

Speed up the uptake of Grid computing within the user community

To motivate the users to join and use the Grid infrastructures, users need easy and reliable access to the computing resources and data stored in multinational Grid infrastructures. It is very encouraging that a large majority of the users are willing to share knowledge, tools, data and results as proposed by the Grid idea.

7.3.2 Future Challenges for HPC Centres

In the year 2001, the beginning of the ENACTS project, the Grid development was in early stages. From 2001, Grid technologies have developed from early pilot projects to working Grid solutions to scientific sub-communities. However, the current Grid infrastructures are still experimental and for the majority of HPC users they are too complicated and time-consuming to use. The Grid development and adaptation by users of HPC centres has been slower than estimated in the year 2001.

There is a clear need for HPC centres to broaden their scope in educating and training researchers in Grid techniques and applications. HPC centres can promote best practice, for example code portability, data interoperability, and Grid enabled applications. HPC centres have already provided support for local Grids and that is important in the future, too. HPC centres should provide appropriate training and education, providing a link between Grid developers and scientific user communities. Grid user support, e.g. helpdesk services to Grid users, end-user documentation and usability of Grid solutions should be of high quality. HPC centres can become the backbone of a true Grid infrastructure offering the scientific community an easy and sound access to HPC and datastores.

The main challenges Grid computing will face in future will be political. The Grid development this far has been very technology-driven. The Grid technologies still need to be developed further, but scientifically and technologically Grids are very possible in the near future. In order to have working Grid solutions, the main issue is how to make political agreements between nations in Europe. More agreements will be needed how to use the data and resources in multinational Grid infrastructures. This requires much closer cooperation between HPC centres in Europe, and also at the international level.

Appendix 1 The questionnaire

Part 1/4

Introduction

The questionnaire is divided into four parts. Please answer the questions below, and press the Proceed button at the bottom to continue to the next section. If you are unfamiliar with a question, you may leave it unanswered.

1. Personal details

Email

Country where you are working

Type of employer

if Other then please specify:

Position

if Other then please specify:

Scientific Profile

2. What is the composition of your group?

☐ Only people from my institution

☐ People from several institutions within my country

☐ International collaborative group but within Europe

☐ International including transatlantic collaboration

3. Group size:

4. Please specify the field of work of your group

5. Is your work multidisciplinary? If yes, please specify your secondary field of work

6. Please give a brief description of your research interests

User Profile

7. Looking at the following definitions, select the approximate percentage that best describes your usage, totalling 100%

I run pre-compiled applications
You specify input parameters and extract results, but you do not modify source code.

I modify the source code when needed
You have access to the source code of your application and you make occasional modifications to the source and/or re-compile the code to use a different set of modules. You do not develop the code.

I am a code developer
You are developing your code as well as using it. Performing new runs often involves re-compilation or modification of the original source.

8. Using the same definitions, select the approximate percentage that best describes your group, totalling 100%

We run pre-compiled applications
Your group specifies input parameters and extract results, but you do not modify source code.

We modify the source code when needed
You have access to the source code of your application and you make occasional modifications to the source and/or re-compile the code to use a different set of modules. You do not develop the code.

We are code developers
You are developing your code as well as using it. Performing new runs often involves re-compilation or modification of the original source.

Proceed to Part 2

Part 2/4

Computing Environment Needs

1. What kind of operating systems do you currently use (tick all that apply) ?

☐ Unix
☐ Linux
☐ Windows
☐ MacOS
☐ Other, please specify:

2. My network connection is

☐ Local Area Network (e.g. University network)
☐ wireless LAN (WLAN)
☐ ADSL (XDSL)
☐ modem or ISDN

3. I connect to my computing environment via

☐ ssh
☐ X windowing
☐ www
☐ client on my own machine

4. What is your primary access to your computing environment ?

☐ Access locally inside my home institution
☐ Access outside my own institution but within my home country
☐ Access to abroad, please describe:

5. Connections (bandwidth) from your own machine to your computing environment

6. Quality of the user support

poor ○ ○ ○ ○ ○ good

7. Information about service breaks (maintenance, updates etc.)

poor ○ ○ ○ ○ ○ good

8. Usability of computers (hardware/software failures and service breaks)

poor ○ ○ ○ ○ ○ good

9. Completion time of batch jobs

too long ○ ○ ○ ○ ○ satisfactory

10. Response time in interactive use

too long ○ ○ ○ ○ ○ satisfactory

11. How do you use your current computing environment ? Please select the percentage that describes your usage in terms of CPU hours, totalling 100%

Parallel computing
Serial computing
Other, please specify:

12. How would you prioritize the following features of the computational resources ?

Speed of single CPU (computational time)	not important	○	○	○	○	○	important
Communication between CPUs (speed)	not important	○	○	○	○	○	important
Memory per CPU (problem size)	not important	○	○	○	○	○	important
Reading/Writing data on disks (I/O time)	not important	○	○	○	○	○	important
Disk space (data size)	not important	○	○	○	○	○	important
Serial or cluster setup	not important	○	○	○	○	○	important
Parallel setup scaling up to 8 CPUs	not important	○	○	○	○	○	important
Parallel setup scaling up to 32 CPUs	not important	○	○	○	○	○	important
Parallel setup scaling up to 128 CPUs	not important	○	○	○	○	○	important
Parallel setup scaling up to 512 CPUs and over	not important	○	○	○	○	○	important

13. For your application, which are the bottlenecks in your current computing environment ?

- ☐ Speed of single CPU (computational time)
- ☐ Communication between CPUs (speed)
- ☐ Memory size (problem size)
- ☐ Reading/Writing data on disks (I/O time)
- ☐ Disk space (data size)
- ☐ Serial or cluster setup
- ☐ Parallel setup scaling up to 8 CPUs
- ☐ Parallel setup scaling up to 32 CPUs
- ☐ Parallel setup scaling up to 128 CPUs
- ☐ Parallel setup scaling up to 512 CPUs and over

14. How will your needs have changed by the year 2007 with respect to your computing environment ?

Proceed to Part 3

Part 3/4

Application Needs

1. **What is the typical usage of your computing systems, totalling 100%?**

▾ Computing intensive
▾ Data intensive
▾ Other (specify): ☐

2. **The HPC applications that you usually run on your computing systems are (totalling 100%)**

▾ Commercial
▾ Freeware
▾ Open source
▾ Self made codes

3. **What language is your application written in (if more than one please select accordingly)**

☐ Fortran
☐ Fortran 90/95
☐ C
☐ C++
☐ Java
☐ Other (specify): ☐

4. **Rank your available programming environment tools**

Editors	bad ○ ○ ○ ○ ○ good
Compilers	bad ○ ○ ○ ○ ○ good
Scientific libraries	bad ○ ○ ○ ○ ○ good
Performance analysis tools	bad ○ ○ ○ ○ ○ good
Debuggers	bad ○ ○ ○ ○ ○ good

5. What kind of compiler do you use ?

☐ Gnu
☐ Intel
☐ Portland Group (PGI)
☐ Microsoft
☐ Microsoft Visual Studio
☐ Platform proprietary (Aix, Irix, HP...)
☐ Other (specify) ☐

6. Do your applications require specific scientific libraries ?

☐ No
☐ Lapack-Scalapack
☐ BLAS
☐ NAG
☐ IMSL
☐ FFTW
☐ GSL
☐ PetSC
☐ Other (specify): ☐

7. Do your applications read/write data files in the following formats ?

☐ Raw binary
☐ ASCII
☐ HDF
☐ CGNS
☐ FITS
☐ Other (specify): ☐

8. How do you manage your data ?

☐ Database, Oracle
☐ Database, DB2
☐ Database, MySQL
☐ Other (specify): ☐

9. Which of these tasks are important for your HPC work ?

Code development, adding new features	not important ○ ○ ○ ○ ○ important
Porting on new computing platforms	not important ○ ○ ○ ○ ○ important
Codes optimization	not important ○ ○ ○ ○ ○ important
Running applications	not important ○ ○ ○ ○ ○ important
Data management	not important ○ ○ ○ ○ ○ important
Sharing of data for collaborative work	not important ○ ○ ○ ○ ○ important
Data analysis	not important ○ ○ ○ ○ ○ important
Visualization	not important ○ ○ ○ ○ ○ important

10. Can your typical applications run on the following parallel architectures ?

☐ Yes - Message Passing - MPI
☐ Yes - Message Passing - MPI 2
☐ Yes - Message Passing - PVM
☐ Yes - Message Passing - Shmem
☐ Yes - Message Passing - LAPI
☐ Yes - Shared Memory - Open MP
☐ Yes - Shared Memory - Pthreads
☐ Yes - Mixed - MPI + Open MP
☐ No, but I could take advantage of having a parallel version of the code
☐ No, the serial code is sufficient to my needs

11. Would you be willing to spend some time porting your applications to a distributed computing system, i.e. Grid environment ?

☐ Yes, but I need some support

☐ Yes, in near future

☐ No, too difficult due to licensing reasons

☐ No, I do not have the resources

☐ No, I don't see any benefit

☐ My application already runs on Grid environment. If so, please specify:

Proceed to Part 4

Part 4/4

Grid Environment Experience

1. What are your needs for a distributed computing system, i.e. Grid infrastructure ?

More computing power	not important ☐ ☐ ☐ ☐ ☐ important
Grid interface to access tightly coupled architecture	not important ☐ ☐ ☐ ☐ ☐ important
Grid interface to access shared memory architecture	not important ☐ ☐ ☐ ☐ ☐ important
To store my data in distributed databases	not important ☐ ☐ ☐ ☐ ☐ important
Collaborative work with scientists at several locations	not important ☐ ☐ ☐ ☐ ☐ important
Distance learning and training	not important ☐ ☐ ☐ ☐ ☐ important
Other needs, please specify:	

2. Have you attended a web-workshop or distributed collaborative working session ?

☐ Yes, once

☐ Yes, 2-5 times

☐ Yes, 6-10 times

☐ Yes, >10 times

☐ No, but I am interested in experiencing it

☐ No, I think it is useless

3. Have you used the following Grid-based applications ?

☐ Yes, I have participated in an Access Grid event

☐ Yes, I have been in a Virtual Observatory

☐ Yes, I have used the Cactus environment

☐ No

4. Have you already experienced the Grid middleware environment ?

Globus	not heard	☐	☐	☐	☐	☐	experienced
Unicore	not heard	☐	☐	☐	☐	☐	experienced
Avaki	not heard	☐	☐	☐	☐	☐	experienced
Condor	not heard	☐	☐	☐	☐	☐	experienced
NorduGrid	not heard	☐	☐	☐	☐	☐	experienced

Other (specify):

5. Which of these features do you think are critical for a Grid-based distributed system ?

Web-based access (Web portal)	not important	☐	☐	☐	☐	☐	important
Single image distributed filesystem	not important	☐	☐	☐	☐	☐	important
Security	not important	☐	☐	☐	☐	☐	important
High interconnection network bandwidth	not important	☐	☐	☐	☐	☐	important
Efficient schedulers	not important	☐	☐	☐	☐	☐	important
Large aggregate computing power	not important	☐	☐	☐	☐	☐	important
Large aggregate data storage capacity	not important	☐	☐	☐	☐	☐	important

6. Which are the main difficulties in your present experience of the Grid infrastructure ?

☐ Unfriendly interfaces
☐ Insufficient documentation
☐ Low bandwith networks
☐ Too few services/available applications
☐ Low security
☐ Difficult code porting/lack of specific API
☐ Low performances

☐ Difficulty in obtaining trusted Grid certificates

Other (specify):

7. **Are you involved in projects that deal with the development of a Grid infrastructure ?**

, Please give a short description:

Future needs

8. **Which of the following Grid services could your institution host in future ?**

☐ It could provide computing power
☐ It could provide data storage
☐ It could host a node for collaborative work/net-meetings, workshops
☐ No, my institution gets the Grid services from a different institution
☐ Nothing
☐ Other (specify):

9. **Would you be interested in sharing your codes, in order to be used over the Grid ?**

☐ Yes
☐ Yes, but I need to port it on a Grid environment
☐ No, I use proprietary/licenced software
☐ No

10. **Would you be interested in sharing your data, in order to be used over the Grid ?**

☐ Yes
☐ Yes, but with specific conditions
☐ No

11. **Do you think that sharing your code/data can increase the visibility of your work/results ?**

○ Yes, because

○ No, because

12. **Before finishing are there any other comments, ideas, suggestions about HPC, Grid and their possible developments?**

Submit the Last Part