A Grid Computing Centre at Forschungszentrum Karlsruhe

Response on the Requirements for a Regional Data and Computing Centre in Germany[1] (RDCCG)

5. December 2001

H. Marten, K.-P. Mickel, R. Kupsch

Forschungszentrum Karlsruhe GmbH
Central Information and Communication Technologies Department, HIK
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen

[1] Requirements for a Regional Data and Computing Centre in Germany (RDCCG), P.Malzacher et al., July 1, 2001

Introduction

Forschungszentrum Karlsruhe (FZK) as a member of the Helmholtz Gemeinschaft Deutscher Forschungszentren (HGF) intends to establish and operate a GRID Computing Centre (G2C) within its Central Information and Communication Technologies Department (HIK) as one of their long term research and infrastructure programmes within the framework of the HGF research areas. This decision has been triggered by the demands of the German particle physics community in view of the unprecedented requirements on data handling and computing arising from the planned experiments at the Large Hadron Collider at CERN (LHC) but it is expected to have strong impact also on other science areas. Such a programme fits perfectly to the mission of the Helmholtz Gemeinschaft focussing on long term research areas of general interest for the public and the society particularly providing a research infrastructure for a widespread national and international users community.
As the first step in this programme at FZK a Tier B cluster for the German participants of the BaBar experiments at SLAC has been installed and is operating safely serving about 25 users from Bochum, Dresden and Rostock. Currently a testbed is set up to participate in the data challenges of the ALICE experiment starting later this year. Intense planning on the set-up of a LHC-Tier1 centre is going on in closest contact to the CERN LHC Computing Grid Project and other TIER1 centres in Europe simultaneously seeking the involvement of external Grid computing competence in Germany as e.g. at Zuse Institute Berlin ZIB, GSI Darmstadt or DESY Zeuthen.

On 1st of July 2001 following a memorandum from May 2001 the German Particle and Nuclear Physics Community represented by their committees KET and KHK respectively laid down the requirements for a Regional Data and Computing Centre in Germany (RDCCG). The main aim of the RDCCG is to provide a computing environment for the German Particle and Nuclear Physics Communities capable of meeting the computing requirements of the future LHC experiments at CERN. It is proposed to establish the Regional Data and Computing Centre Germany to serve as a Tier 1 centre for the four LHC experiments, thus assuring competitive analysis resources for the German research groups involved. The RDCCG should also serve other major data intensive non-LHC experiments of both communities and should extend to other sciences later on.

Grid computing technologies will be developed in close cooperation with other LHC computing centres and with ongoing and future Grid activities like DataGrid, PPDG and GriPhyN to satisfy the extreme requirements in terms of computing, data storage and networking that the LHC experiments will need. The RDCCG should both have centralised hardware resources and act as a focal point for Grid software development. Advantage should be taken of competence in computer science and Grid computing already available in Germany. The centre should encourage other branches in natural sciences to join activities in this promising field. It should take a leading role in education of young people in applied computing and thus constitute a visible investment into the future of research.

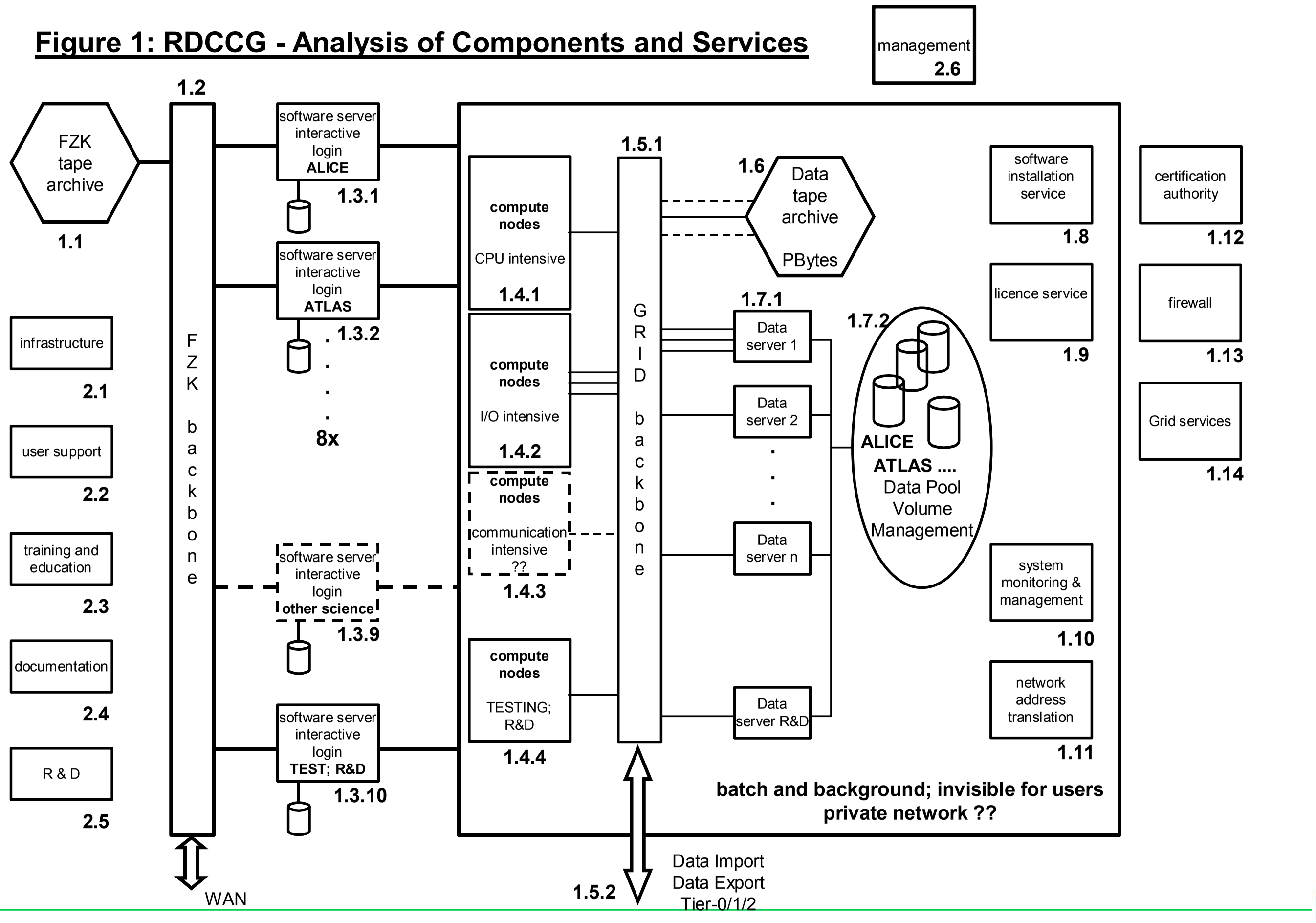
Three stages of development are suggested for the RDCCG:

- A test bed for developing the infrastructure of LHC computing between 2001 and 2004. At this first stage experience could be gained from serving active non-LHC experiments.
- After a complete appraisal of all developments, the years 2005 to 2007 should be used to finalize the computing infrastructure for the LHC experiments.
- From 2007 onwards the centre should adapt to the increasing needs of the ongoing LHC experiments and should remain fully operational for the whole duration of the LHC experiments, at least 15 years.

Forschungszentrum Karlsruhe has been asked to analyse and comment on the requirements for the Regional Data and Computing Centre Germany. This paper represents the response of the FZK Grid project group to these requirements and is thought to serve as a basis for further discussions and cooperation with the users community. Figure 1 summarizes all components and services that were identified in the paper of the German Particle and Nuclear Physics Communities. This Figure should give a simple overview over most of the necessary components of RDCCG and their possible integration into the infrastructure of Forschungszentrum Karlsruhe, and it may be used as a guideline through the following discussions.

Forschungszentrum Karlsruhe
Technik und Umwelt

Figure 1: RDCCG - Analysis of Components and Services



1 Hardware Requirements

1.1 FZK Tape Archive

FZK currently operates two Powderhorn Silos 9310 with a total capacity of 12.000 cassettes with 20 GByte each (STK 9840), giving a total tape storage capacity of 240 TByte native. The Tivoli Storage Manager, TSM, is used to serve all current requirements for data backup and archive of the whole Research Centre. Some tape drives will be replaced at the end of 2001 by STK 9940B (200 GByte native per cassette in 2002) and integrated into a Storage Area Network (SAN).

For resource sharing and synergy effects it is recommended to use these systems as well for the backup of experiment specific software on the servers 1.3.x (see Section 1.3 below) and for the backup of user-specific data ($HOME) on these systems. The estimated data volumes to be transferred to tape are:

in 2002: (10 users per experiment x 1 GByte $HOME + 50 GByte experiment specific software)
x 10 server x safety factor 5 = 3 TByte

in 2007: (100 users per experiment x 1 GByte $HOME + 50 GByte experiment specific software)
x 10 server x safety factor 5 = 7.5 TByte

It is not recommended to run backups for massive temporary data from /tmp, /scratch, /work,..., or to archive user data. The latter should be done by every user at his home institution.

1.2 FZK Backbone

The FZK backbone consists of a 1 Gbit LAN infrastructure with a 128 Gbit/s back plane, daily used by about 3000 scientists of Forschungszentrum Karlsruhe and of the University Karlsruhe. A WAN connection with 34 Mbit/s for a data transfer volume up to 1280 GByte/month is currently available via the German Research Network, DFN.

It is envisioned to use the FZK backbone for the backup of the experiment specific software server 1.3.x as well as for the interactive login of LHC- and non-LHC users at the start-up of the RDCCG. However, it is not planned to use the FZK backbone for massive transfer of experiment specific data, except during a possible start-up phase, i.e. as long as no additional WAN connections for Grid computing are available. Massive data transfers over the FZK backbone would need a complete redesign of the whole networking infrastructure, including internet routers, firewall components etc., together with respective installations in a complex running system. Estimates for required hardware and manpower investments indicate that it is easier and less expensive to build a new Grid backbone infrastructure (Section 1.5.1) with the ability to adapt to the increasing needs and fast changing requirements during the next 4-7 years.

1.3 Experiment Specific Software Server

The experiments ask for special administration areas and rights to install their experiment specific software by own experts. Furthermore, they wish to have the possibility for interactive login.

An analysis of the working methods of the HEP experiments has shown different requirements on the installed operating systems, development environments, software packages, licensed software etc., down to the level of specific OS kernel patches and software releases. Additionally, we identified collaborations between the HEP experiments at different levels of software developments on the one hand, but numerous scientific competitions on the other hand. While collaborations need common development environments, competition requires a separation of software areas and special data protection mechanisms.

To separate especially the different underlying operating systems, software releases and security requirements of the HEP experiments, we propose to install (for “fail save” operation at least) one dedicated software server per experiment (1.3.1-1.3.8 in Fig. 1). These servers could be used for experiment specific software installations and as the desired reference systems for software tests. Furthermore, they may serve for home directories of single users at RDCCG start-up. Installed software packages as well as home directories should be backed-up into the FZK tape archive (1.1) via the FZK backbone. Additionally, these servers could be used for the desired interactive login at RDCCG start-up. However, in the sense of Grid computing it should be envisioned that interactive work only remains necessary for system and software administrators on a midterm time scale. User specific software and data should then be stored at the respective home institutions.

The experiments wish to install the experiment specific software on these servers by own experts. For such cases, the terms of reference for data protection and security at Forschungszentrum Karlsruhe demand that one person per experiment (i.e. for each server) takes the full responsibility for these installations. At least two experts per experiment should be named for the installation of the experiment specific software. The same people should be responsible for the import and export of experiment specific data and should be available for experiment specific user support. It is very likely, that RDCCG personnel cannot solve massive problems with experiment specific software packages without the support by respective experts. The installation of any software package should be well documented in a universal documentation system (2.4) to enable a detailed error analysis in case of failures.

The provision and first installation of these software servers will be done in close collaboration between the RDCCG and the experiments. In order to save money and manpower it is important that the experiments give a detailed list of requirements for the hardware as well as for the software and licences running on these systems. The initial hardware investment at RDCCG start-up for these servers largely depends on the size, quality and design for fail safe operation. One of these servers is already installed for BaBar; the next two systems will be purchased as soon as the hardware requirements have been specified.

Additional software servers might be installed for other sciences and for testing, research and development (1.3.9 and 1.3.10 in Fig. 1).

1.4 Compute Nodes

A large number of compute nodes is necessary to satisfy the massive requirements for compute power of the HEP experiments. Currently, Linux running on commodity “Intel-like” dual processor boards are the preferred systems.

The current average lifetime of processor boards on the commodity market is only about 6-12 months. Thus, there is no doubt that a compute cluster of the envisioned size, which is developed, upgraded and maintained over many years, will necessarily be inhomogeneous. Care should be taken that the user software is as independent as possible of the hardware, underlying operation systems, compilers and vendor specific libraries.

It is planned to run the whole system in batch mode only. This allows a flexible resource management and resource sharing between all HEP experiments via corresponding queuing systems. Open source products are preferred against commercial systems because of tremendous licence costs for clusters with hundreds or even thousands of nodes and for hundreds or thousands of users, and because of the flexibility to fit these tools to local requirements. Respective products are under development in numerous international projects, not only for the local cluster management but also for Grid infrastructures with interfaces to the local systems.

Although the current planning assumes “commodity systems”, a few basic requirements should be fulfilled for a cluster to be managed and permanently upgraded to a few thousand processors over many years:

- Rack based compute nodes to save floor space and allow easy upgrades or exchanges of entire cluster sections (supplying of complete, tested racks by vendors).
- Service friendly systems which allow an easy exchange of system disks
- Temperature and fan control as well as sound internal cooling systems for fool proofness
- Low power consumption i.e. low heat production
- As few components per system as possible

It is worth to notice that in this sense “commodity” is not equivalent to “low cost”. The certainly higher investment for such systems is returned by respective savings e.g. for personnel, electricity or main cooling systems.

Further costs might be saved if sections of the cluster are designed for special purposes instead of trying to design one huge general-purpose system. For example, typical Monte Carlo simulations in high energy physics are certainly CPU bound and do not necessarily require expensive high speed interfaces to mass storage and fast switches, in contrast e.g. to typical filter jobs running over large data bases. Another type of codes are real parallel applications that might need special high-speed networks for inter-process communication. For these reasons, the compute nodes in Fig. 1 are divided into different blocks for CPU intensive (1.4.1), I/O intensive (1.4.2) or communication intensive (1.4.3) tasks. Respective special requirements and costs should be analysed repetitively during the cluster upgrade within the next years. The cost estimates presented in the Appendix are based on “simple” general purpose systems and do not take into account such possible special needs. Also, reference systems with other processors than “Intel-like” and with alternative operating systems are currently not planned and are therefore not taken into account in the cost estimates.

1.5 Network

1.5.1 Grid Backbone (LAN)

The Grid backbone is the central component of the whole computing environment and should be designed for fail save operation. An upgradeable Gigabit switch with a 128 Gbit/s back-plane, redundant management modules and redundant power supplies will be available at the beginning of 2002. Management software present at Forschungszentrum Karlsruhe will allow to measure and record transfer rates down to single ports as well as the overall LAN throughput from the beginning, in order to get a better insight into the technical requirements for future upgrades.

1.5.2 Grid WAN Connection

Dedicated WAN connections to other Grid computing centres are necessary for data import and export. One of the GWin nodes (622 Mbit/s) of the German Research Network, DFN, is located in the building of the FZK Computing Centre, at a distance of a few metres to the Grid backbone (1.5.1). However, the cost for high-speed/high-volume data transfer is as high as the risk for false investment if the respective partners cannot deliver or receive data at the same data rates. Therefore, it is recommended to start the discussions between respective Tier 0/1/2 centres on the current needs as soon as possible. It is likely that new dedicated connections for data import/export need special high-speed routers and firewall techniques which demand for large initial investments. Since Grid security mechanisms and policies are not yet well defined, the cost for such systems can only be seriously estimated after the requirements have been well defined in collaboration with other centres.

1.6 Data Tape Archive

The data tape archive (1.6 in Fig. 1) is planned as the background storage element for data on a Petabyte scale. The roadmaps of different vendors for almost 1 TByte per tape and 100 MByte/s per tape drive within the next 4-5 years are quite promising. Today, state-of-the-art are scalable systems up to 11 PByte with 200 GByte-tapes.

Studying the roadmaps of two manufacturers gives a rough idea of the future innovation:

Roadmap of manufacturer A (data not compressed):
2001: 200 GByte/Tape; 24 MByte/s drive
2002: 400 GByte/Tape; 48 MByte/s
2005: 800 GByte/Tape; 96 MByte/s

Roadmap of manufacturer B (data not compressed):
2001: 100 GByte/Tape; 15 MByte/s drive
"Factor 2 in tape packing density each year"

Tape qualities and prices largely depend on the type of access. The planned tape access is of the type “write once, read very rare”.

Also important are information on the required data rates, band widths and access times, which determine the necessary number of drives and therefore the price of the whole system. Another open question is the type of the required management software. To implement the hardware and software compatibility with CASTOR we need support from an experienced person.

Finally, in the area of tape archives two things should be always kept in mind:

- Initial investments for a robot system and drives are very high
- Changing a technology is manpower and cost intensive

1.7 Online Data Management

1.7.1 Data Server

A step towards a “fail save” and scalable high performance solution is clustering of file servers and respective data management. These clustered file servers should be available for all experiments (synergy). The management of disk pools through an adequate Volume Management Software is an excellent tool to reduce capacity shortages of one user group while removing over capacities of another group at the same time. Such Volume Managers are available e.g. under Linux. To optimise the performance of the system the overall band width and workload for each experiment is needed. Our particular interest is information on required transfer rates from CPU to disk, disk to tape and tape to disk within the computing centre as well as between different computing centres in the Grid.

1.7.2 Data Pool

To avoid frequent failures it is advantageous to use the largest available disk capacities, i.e. to keep the number of components as small as possible. Some steps towards "fail save" are Raid-5 systems, redundant controllers and power supplies, uninterrupted power supply units and hot swap devices. Advanced controllers (e.g. Fibre Channel for SAN) are essential to meet massive data transfer rates in the Grid. For practical reasons rack-based systems are preferred to single components. Also important are continuous monitoring of the components and extended (36 months) warranties since especially disk arrays do not always work with a mixture of components.
The overall installation and management of the data pool will be carried out by RDCCG, import and export of data is carried out by the experiments themselves.

The RDCCG strongly prefers to use SCSI-disks, because a high rate of failure of IDE-disks was observed in many different systems in recent years. A high rate of failure needs additional manpower and therefore an extra amount of fixed costs.

1.8 Software Installation Service

To support a big number of processors it is indispensable to have an automatic installation service for operating systems as well as for non-experiment specific software. FZK has gained good experience with central Linux installation services and software distribution mechanisms. Two redundant servers are recommended to obtain good availability, however redundancy of this service is not stringent at RDCCG start-up.

While experiment specific software environments on the respective servers 1.3.x do not impose larger problems, a homogeneous software environment is recommended on the compute nodes 1.4. A complete catalogue of non-experiment specific software which should be available in 2002 is urgently needed.

1.9 Licence Server

For the management of licences three redundant servers are obligatory, at least two of them must be permanently available to check for licence keys. These servers do exist in the computing centre of

FZK, and it is planned to replace them in 2002 because of their advanced age. Probably these systems can be used for Grid computing as well.

1.10 System Monitoring and System Management

A good system monitoring and management tool is indispensable for reliable operation of a computing centre of the envisioned size (e.g. one hard disk failure per day is expected in a system with 1800 disks). FZK has gained wide knowledge and experience with Tivoli, which is used since many years for central system monitoring and management. However, the licence conditions for hundreds or thousands of processors in a local Grid environment are still an open question. On the other hand there is a good chance of using the tools from current Grid projects like DataGrid in the near future, and to participate in the development and testing of these tools by a new division devoted to Grid computing research and development at HIK (see also Section 2.5). Still, it is worth to notice that the handling of such a system management tool is manpower intensive even if many things can be automatized. In the case of a commercial product it is also cost intensive.

1.11 Network Address Translation

The question of whether network address translation (NAT) is necessary is directly connected to security issues and possible difficulties to manage a new RDCCG network within the already existing one. Forschungszentrum Karlsruhe was among the first organizations having an own Internet access with about 64.000 adresses in a public class A network. For the RDCCG, of the order of 10.000 new components have to be integrated during the next years. These would all be visible to the general public, with respective implications on the overall security. Also, an inevitable spreading of the respective new addresses into discontinuous address ranges would be rather error-prone. For these reasons of safety, reliability and flexibility RDCCG will start operating in a private network with a restricted number of links to the outside. NAT then enables a batch-job to communicate with external addresses. First tests under Linux have shown positive results. However, as mentioned many times in this paper, sound ideas to solve security issues on the Grid are still lacking. Whether NAT is the preferred method to solve these security and management issues must be elaborated in close collaboration with other Tier 0/1/2 centres in future projects.

1.12 Certification Authority

FZK has acquired wide knowledge on building an own Certification Authority (CA). The PGP-software, respective hardware as well as infrastructure to prevent access by non-authorized persons is already ordered. This service will be available at the beginning of 2002.

The extension to Grid services is essential, but does not raise general new difficulties. Small additional manpower is necessary to adapt the system to the special needs on the Grid (e.g. OpenSSL instead of PGP) and to install a national Grid CA.

1.13 Firewall

Firewall techniques are implemented at Forschungszentrum Karlsruhe to avert a permanently increasing number of attacks from the outside. Respective hard- and software as well as know-how are available. To our knowledge, Grid security aspects are not yet fully understood, but they will become important as soon as the RDCCG will be feasible. Thus, firewalls and other security mechanisms must be discussed as soon as different Grid computing centres are linked together in a non-private network. Manpower and money should be reserved for these issues already in 2002.

1.14 Grid Services

Many of the required Grid services, like e.g. those for automatic data replication between Grid Computing Centres, are still under development in numerous international Grid projects and are therefore not ready up to now. The RDCCG will either participate in these developments or will gain experiences by testing prototypes (see also Section 2.5 on Research and Development).

2 Infrastructure

The following sections summarize the infrastructure necessary to operate the RDCCG. This infrastructure comprises technical, like e.g. electricity or air-conditioning, as well as managerial or educational issues, like the general management of RDCCG, user services, on-call duties or education and training.

2.1 Infrastructure (room planning, electricity, air-conditioning)

The size of the suggested computing centre becomes quite obvious when looking at the necessary technical infrastructure. More than about 300 squared metres of floor space, about 150x 380 V and 300x 230 V sockets and more than 600 kW air-conditioning are needed to build up the hardware of the RDCCG in the final extension. These are real cost factors because not only the equipment has to be paid, but also the installations which have to be carried out necessarily by external companies. Up to now no complete cost estimation can be given. The realisation needs some time in advance for planning, submission, placing, carrying out and so on. Estimates on the cost of operation are given in the Appendix.

2.2 User support and User Administration

The FZK Computing centre operates a central hotline & helpdesk support for all services since about two years. The user administration takes place via central, semi-automatic registration and logoff procedures in coordination with a contact person of each scientific institute.

Similar structures will be implemented for LHC and non-LHC experiments. Consequently the naming of a contact person for each experiment as a coordinator is required. The RDCCG is responsible for user support in the area of OS (Linux), non-specific experiment software, compiler and programming languages as well as for the general Grid-software (RDCCG and Competence Centre). New services for Grid will be implemented for directory services, helpdesk structures etc.

User support for experiment specific software and data storage has to be done by the experiments.

2.3 Education and Training

FZK has an own division for education, the so-called Education Centre for Technology and Environment (FTU). Via FTU, special courses on operating systems, computing languages and software tools are already offered today, and additional lectures e.g. on Grid computing software may be easily added. Additionally, an in-house training with good access to the Grid resources may be arranged. Furthermore, different institutions of the University of Karlsruhe already signalised strong interest to collaborate in the field of Grid computing, and common lectures for students could be established additionally to the already existing common meetings on Linux-Clusters. These activities will be a main part of the Grid Competence Centre with support by RDGGC personnel.

2.4 Documentation Management

The FZK computing centre currently uses different tools for documentation management. Most important are frequently updated web pages. For problem management a helpdesk system based on the action request system of the Peregrine (formerly Remedy) Corporation is installed. For general questions a news server is available. In 2002 a web based interface for the documentation of installed hardware and software will be developed, and a helpdesk system will be adapted to the special needs of Grid-users. The RDCCG is responsible for the documentation of OS, non experiment-specific software and a description of the access to Grid components. Experiment-specific software should be documented by the experiments.

2.5 Research & Development (Grid Competence Centre)

Embraced by the term “Scientific Computing”, Grid-Computing will be one of the key technologies in the FZK research program from 2002 onwards: The newly founded Grid-Computing division will start working on the 1st of January 2002 and establish a Grid Competence Centre; the head of the division is already promoted. The Grid Competence Centre will focus on software development in three strategic areas of concern:

- Grid Fabric
- Grid Middleware
- Grid Applications

The basic orientation of the new division is to collaborate with all sciences on- and off-site, with HEP and informatics as prime partners, but it will also cooperate with industrial and commercial partners. The competence centre will support RDCCG in all aspects relevant to Grid computing and in the participation in LHC data challenges. In the field of fabric management, the operation of large fault-tolerant computer clusters as well as the analysis of data flow in the system and the realization of storage concepts will be a major topic. The Grid middleware will cover not only the further deployment of Globus, but in addition study the use of object oriented new approaches like .NET. In the area of application development, FZK is a main contractor of the EU CrossGrid project, and the new division will take over responsibilities in the work package management of the Grid application program environment. CrossGrid has a strong commitment to application development based on the DataGrid software releases; the integration of the corresponding HEP applications will take place at the testbed site in FZK. Furthermore, FZK is a member of the international “LHC Computing Grid Project” to be launched at the end of 2001, and will support the linking of CrossGrid developments with these activities.

2.6 Management

Forschungszentrum Karlsruhe supports the suggestion to establish an Overview Board (OB) and a Technical Advisory Board (TAB).

The OB decides on general policy, resource and funding issues affecting the RDCCG and arbitrates in case of conflicts on technical issues. The membership of the OB should consist of:

- A member of the FZK board as chairman,
- Computing director of the centre,
- Project Leader (PL) and deputy,
- A member from BMBF,
- Respectively one member from KET and one from KHK,
- Two members from LHC experiments and two members from non-LHC experiments (i.e. four persons in total),
- chair of the TAB.

Every named person should have a substitute. The OB should meet at least once per year, or more frequently if required, to review and agree on major changes in human and material resources and the project plan.

The TAB consists of:

- One software and computing expert of each experiment, which is also nominated as dedicated technical contact person to the RDCCG,
- One representative from the CERN “LHC Computing Grid Project”,
- One representative from each of the three remaining European LHC Tier 1 Centres IN2P3, INFN and RAL,

- PL and deputy,
- One representative of the non-German LHC Tier 2 Centres (to be reduced to two representatives in total later-on)
- Respectively one representative from DESY, from KET and from KHK.

Every named person should have a substitute. The TAB may decide to invite other experts to its meetings. The first meeting is initiated by the centre. At this meeting the internal structures of the board will be established.

The TAB helps to elaborate a detailed work programme, using inputs from the experiments. The PL informs the TAB of all relevant technical issues at every meeting and seeks agreement with the TAB. In case of disagreement the OB has to be involved. The TAB should meet at least two times a year.

The PL is responsible for the day-to-day management of the RDCCG and for the developing of a project plan in close relationship with the TAB and the CERN-SC2, and manages the execution of the approved project plan. He reports to the computing director of the centre and to the OB and is responsible for a close liaison to the physics community.

Each experiment nominates a dedicated technical contact person to the RDCCG (see membership of TAB) and the RDCCG provides a corresponding contact expert.

The design, construction and operation of the RDCCG over the years is a challenge in itself. Very important is applying for sponsorships together with the Grid Competence Centre and the German Particle and Nuclear Physics Community. The integration of additional organizational units of Forschungszentrum Karlsruhe for planning, success control and external relations should be envisioned.

2.7 Operation of the RDCCG

Hardware and software compatibility with respect tot the other centres is very important. The RDCCG suggests the following procedure. The RDCCG (Tier 1) come to an agreement with CERN, the Tier 2 centres come to an agreement with the Tier 1 centre and so on.

The operation of the centre is provided 24 hours a day and 7 days a week. Operators and experts are available during on-site working hours. In hours of non-supervised operation an on-call expert service for urgent technical problems is under consideration. Because of cost optimisation not every component of the system can be set up for fail save operation. Hardware maintenance is ensured, service contracts will be concluded under best price/performance conditions.

3 Hardware Requirements

According to the paper “Requirements for a RDCCG”, the needs for LHC computing are divided into three different phases:

- Phase 1: Development and Prototype Construction, 2001-2004
- Phase 2: Installation of the LHC Computing Production Facility, 2005-2007
- Phase 3: Maintenance and Operation, 2007 until the end of the LHC data analysis

Additionally, four non-LHC experiments ask for resources in a stable production environment in the years 2001-2007. The hardware requirements during the above three phases are summarized in the following subsections. Required resources for BaBar after 2005 have not yet been specified and are thus not taken into account in the following tables.

3.1 Hardware Requirements

3.1.1 Phase 1: Development and Prototype Construction, 2001-2004

During this first phase, milestones have been defined at month/year 11/2001, 4/2002, 4/2003 and 4/2004, at which prototypes of the computing centre should reach a given size. The following table summarizes the additional resources to be installed each year for LHC and non-LHC (nLHC) experiments to reach the prototypes at the respective months. According to these suggestions, shared CPU resources between all experiments are taken into account, while disk and tape capacity are listed separately for LHC and nLHC experiments, respectively.

	month/year	11/2001	4/2002	4/2003	4/2004
LHC + nLHC	CPU /kSI95	1	8	10	27
LHC	Disk /TByte	1	10	23	24
LHC	Tape /TByte	1	20	35	45
nLHC	Disk /TByte	6	28	45	66
nLHC	Tape /TByte	6	84	65	89

3.1.2 Phase 2: Installation of the LHC Computing Production Facility, 2005-2007

The following two tables summarize the required additional resources per year between 2005 and 2007. No resource sharing is foreseen during this second phase, thus resources are given for LHC and nLHC experiments separately.

LHC year	2005	2006	2007
CPU /kSI95	77	172	580
Disk /TByte	109	298	490
Tape /TByte	307	983	1528

nLHC year	2005	2006	2007
CPU /kSI95	11	5	5
Disk /TByte	125	98	98
Tape /TByte	154	210	210

3.1.3 Phase 3: Maintenance and Operation, 2007 – end of LHC data analysis

No required resources are given for non-LHC experiments after 2007. Thus, the following tables lists the final configuration in 2007 for LHC experiments only.

LHC year	2007
CPU /kSI95	829
Disk /TByte	897
Tape /TByte	2818

It is assumed that the capacity of this system is increased after the construction period by a constant amount of the 2007 value every year:

- The amount of CPU is increased every year by 33%. The obsolete equipment is replaced after the 3-year maintenance period.
- The disk space is increased every year by 50%, to make the analysis of bigger data samples possible.
- The tape storage capacity is increased every year by 100%, to store the new data.

3.1.4 Summary of Hardware Requirements during the three Phases

The following table summarizes the total (cumulative) hardware for LHC and nLHC experiments, which should be available at RDCCG after the respective milestones are reached.

month/year LHC+nLHC	11/2001	4/2002	4/2003	4/2004	2005	2006	2007	2007+
CPU /kSI95	1	9	19	46	134	311	896	+276/yr
Disk /TByte	7	45	113	203	437	833	1421	+450/yr
Tape /TByte	7	111	211	345	806	1999	3737	+2818/yr

3.2 Volumes to be purchased

In order to estimate the prime cost of the hardware, three aspects must be taken into account.

At first, installed hardware must be replaced after a certain maintenance period (typically 3 years) so that the volumes to be purchased every year are larger than the pure requirements listed in Section 3.1.4.

At second, the estimated prime cost per year for each of the three components CPU, disk and tape is well above 200.000 Euro (except for prototype 0 at 11/2001). Thus, for every procurement a European-wide competition of vendors is necessary, which typically needs about half a year. Consequently, hardware prices to be taken into account are typically those at about three months before the respective milestone. This backward time shift by 3 months is taken into account in the following tables as well as in the final cost estimates.

Finally, triggered by the recently launched Grid activities of Forschungszentrum Karlsruhe, and because of the above mentioned large time scales for procurements, FZK already made large investments into new hardware. These investments are taken into account in the following tables by defining two new prototypes at 4/2002 and 10/2002. The first one represents those resources which will essentially be available for Grid computing in April 2002. The size of the second prototype at 10/2002 is equal to the original prototype 1 from the RDCCG requirements, i.e., the difference between both new prototypes at 4/2002 and 10/2002 is equal to the necessary additional hardware investment in 2002 to reach prototype 1 of the RDCCG requirements. For the cost estimate of these new prototypes the above mentioned backward time shift by 3 months is taken into account, so that the necessary investments appear at 1/2002 and 7/2002 in the tables below.

3.2.1 Phase 1: Development and Prototype Construction, 2001-2004

No hardware replacements are necessary during this first phase, and CPU, disk and tape systems for LHC and nLHC experiments can be purchased at the same time. Thus, the volumes to be purchased in each year are combined for both kind of experiments in the following table:

month/year LHC+nLHC	1/2002	7/2002	1/2003	1/2004
CPU /kSI95	3	6	10	27
Disk /TByte	14	31	68	90
Tape /TByte	60	51	100	134

3.2.2 Phase 2: Installation of the LHC Computing Production Facility, 2005-2007

During this phase a completely new installation is assumed for the LHC experiments. No replacement of old LHC hardware is taken into account, so that the calculation of the respective volumes to be purchased in each year is straightforward:

LHC year	2005	2006	2007
CPU /kSI95	77	172	580
Disk /TByte	109	298	490
Tape /TByte	307	983	1528

For the non-LHC experiments it is assumed that they continue to use all CPUs as well as their own disk and tape systems from the previous first phase, supplemented by their additional hardware requirements between 2005 and 2007. A replacement of CPUs is assumed after 3 years (i.e. starting in 2005), of disks after 4 years[2] and of tapes after 5 years. These replacements are indicated in braces by "+" in the following table:

nLHC year	2005	2006	2007
CPU /kSI95	11 (+9)	5 (+10)	5 (+27)
Disk /TByte	125	98 (+34)	98 (+45)
Tape /TByte	154	210	210 (+90)

3.2.3 Phase 3: Maintenance and Operation, 2007 – end of LHC data analysis

Requirements for 2007 and beyond are given for LHC-experiments only. The respective volumes to be purchased each year including replacements are given in the following table (until 2010 only):

LHC+nLHC year	2008	2009	2010
CPU/kSI95	276 (+77)	276(+172)	276 (+580)
Disk/TByte	450	450(+109)	450 (+298)
Tape/TByte	2818	2818	2818(+307)

3.2.4 Summary of volumes to be purchased during all Phases

The following table summarizes the volumes to be purchased including hardware replacements during the three different development phases of the computing centre until 2010:

month/year LHC+nLHC	1/2002	7/2002	1/2003	1/2004	2005	2006	2007	2008	2009	2010
CPU /kSI95	3	6	10	27	97	187	612	353	448	856
Disk /TByte	14	31	68	90	234	430	633	450	559	748
Tape /TByte	60	51	100	134	461	1193	1828	2818	2818	3125

4 Integration Plan

The RDCCG set-up of the first prototype at 4/2002 is shown in Figure 2. Available equipment and services will be:

[2] The replacement period of disks largely depends on the quality and reliability. Usually, this it equal to the warranty of the vendors, typically 3 years for IDE and 5 years SCSI disks. In this paper, an average of 4 years is used.

- three experiment specific software server (1.3.x)
- front-end for the EU-project CrossGrid (1.3.x); not available for HEP
- connections to the FZK tape archive (1.1) via FZK LAN (1.2) for backup
- 36 Linux Compute Nodes (dual PIII) for production, i.e. about 3 kSI95 in total (1.4.x)
- 8 Linux Nodes for CrossGrid and internal tests (1.4.4); not available for HEP
- 16 TByte online (net capacity): IDE, FC-SCSI, SAN (1.7)
- 2.4 TByte online (net capacity) for CrossGrid and internal tests (1.7); not available for HEP
- 60 TByte tape (1.6)
- central software installation service (1.8)
- national CA (1.12)
- central Grid (Globus) services (1.14)
- the existing 34 Mbit/s WAN connection of FZK

The most important goals for the first prototype installation are performance measurements as well as classifications of the data on the different systems. No further serious planning can be done without a specification of necessary data rates from CPU to disk, disk to tape and CPU to tape, and of the data classes.

The planned RDCCG set-up of the second prototype at 10/2002 is shown in Figure 3. Additional available equipment and services should be:

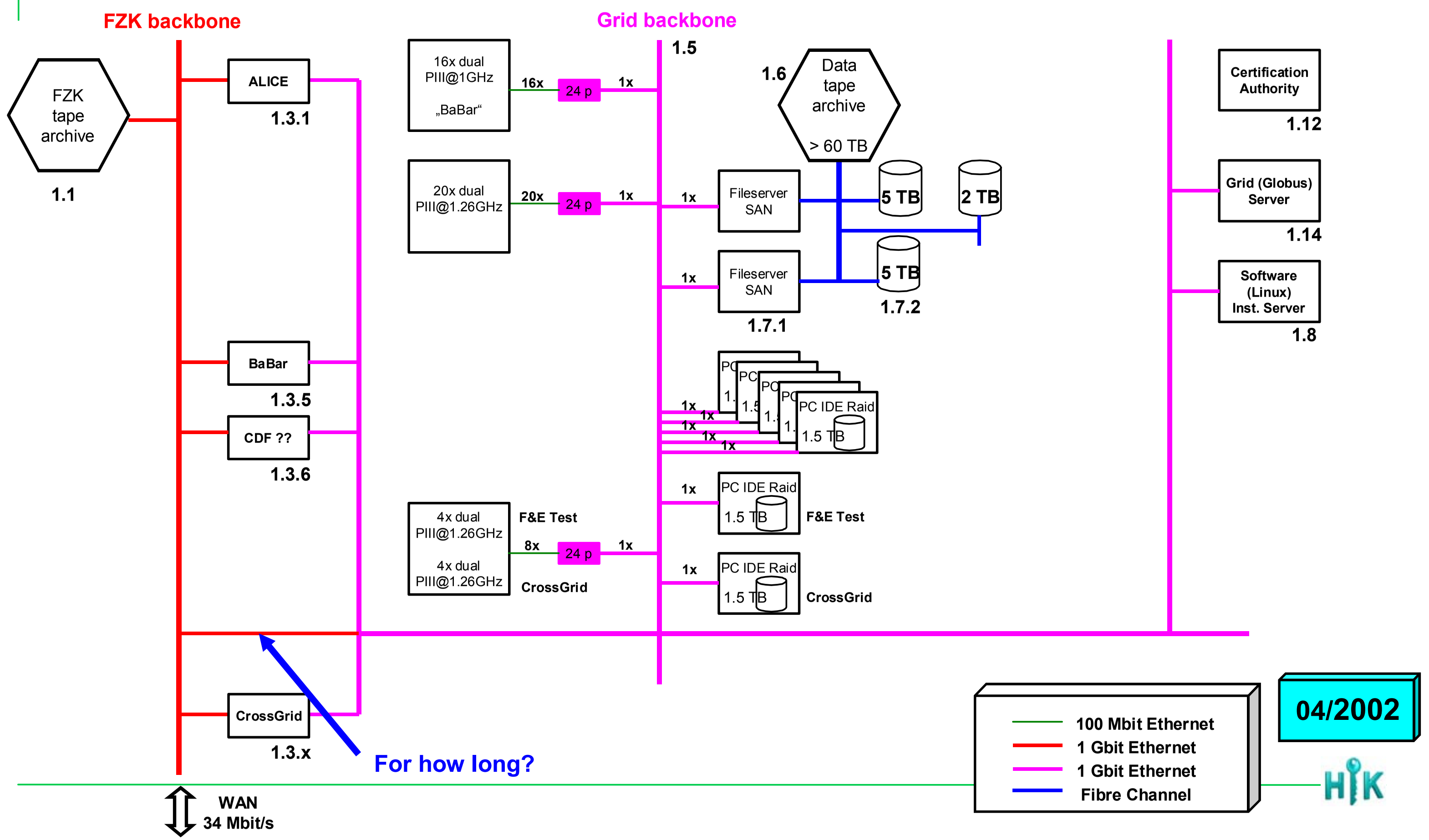
- five additional experiment specific software server (1.3.x)
- + 6 kSI95 CPU (about 160 ? CPUs)
- + 31 TByte net capacity online
- + 51 TByte tape
- dedicated WAN connection for Grid Computing (1.5.2)
- respective security mechanisms, e.g. firewall (1.13) and/or NAT (1.11)

Additionally it is envisioned to have available in 10/2002 (not contained in Figure 3):

- user support with hotline and helpdesk (2.2)
- web based documentation management (2.4)

Forschungszentrum Karlsruhe
Technik und Umwelt

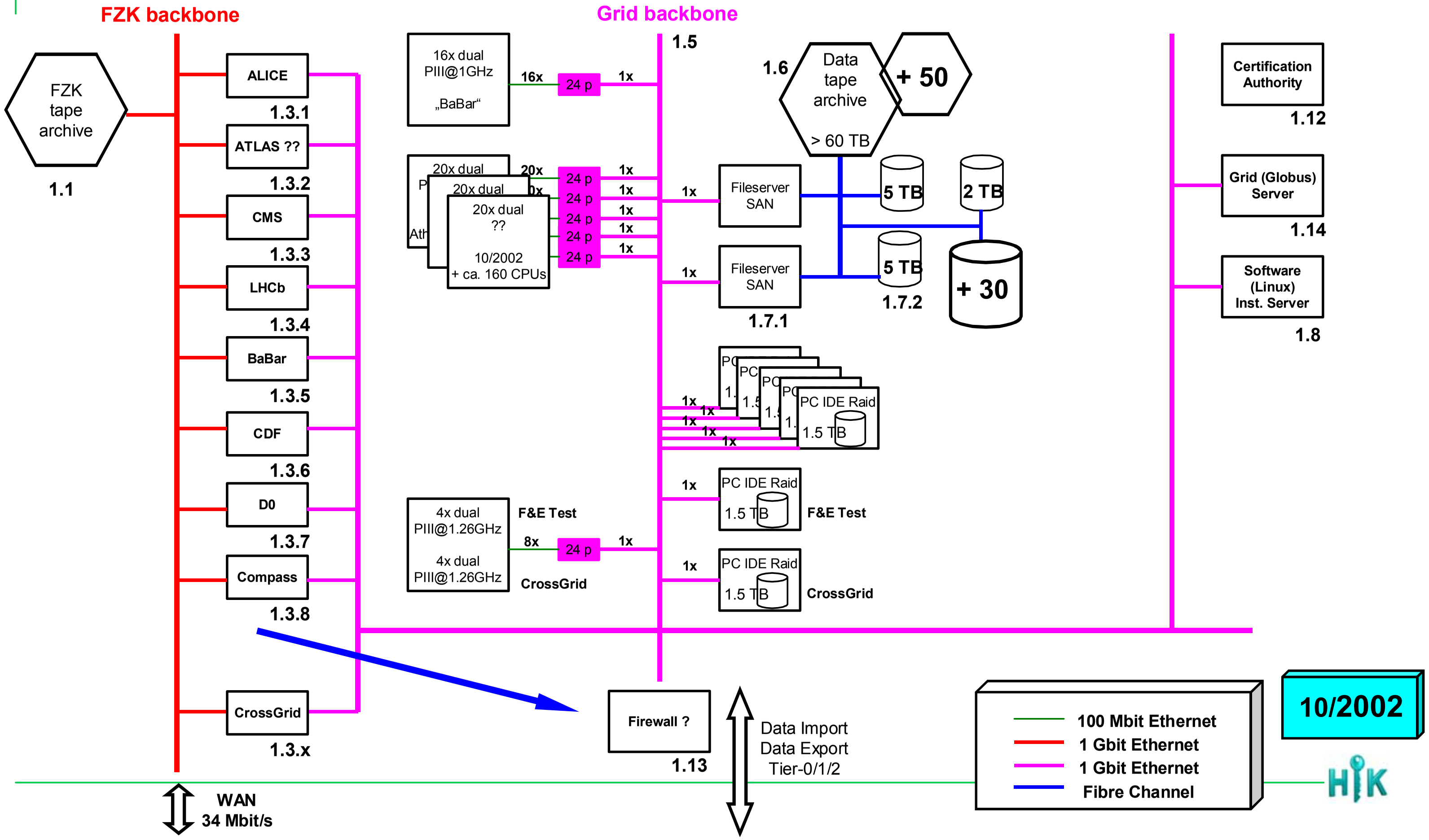
Figure 2: RDCCG - Integration plan 4/2002



04/2002

Forschungszentrum Karlsruhe
Technik und Umwelt

Figure 3: RDCCG - Integration plan 10/2002



Acknowledgements

We appreciate the continuous interest of the Federal Ministry of Education and Research, BMBF. This work was supported by the BMBF under grants 06B0 904I (BaBar) and 06 KA 997I (Hadronen- und Kernphysik).

bmb+f - Förderschwerpunkt
BABAR
Großgeräte der physikalischen Grundlagenforschung

bmb+f - Förderschwerpunkt
Hadronen - und Kernphysik
Großgeräte der physikalischen Grundlagenforschung