The Research Council
of Norway

# Investments in the Norwegian eInfrastructure for Computational Science

## An investment plan for the period 2008-2017

The Research Council of Norway
P.O.Box 2700 St. Hanshaugen
N–0131 OSLO
Telephone: +47 22 03 70 00
Telefax:    +47 22 03 70 01
bibliotek@forskningsradet.no
www.forskningsradet.no/english

The report can be ordered at:
www.forskningsradet.no/publikasjoner

or green number telefax: +47 800 83 001

Printing:  The Research Council of Norway
Number of copies:  110

Oslo, October 2008

ISBN  978-82-12-02594-3 (printed version)
ISBN  978-82-12-02595-0 (pdf)

# Preface

The document contains the advise from the Advisory Committee for Investment in eInfrastructure (ReInfra) to the eVITA Programme Board regarding long term investment strategies for eInfrastructure (high performance computing, storage and GRIDs) in Norway. Three suggested investment strategies are presented based on three budget scenarios: low budget – 22 MNOK/year, medium scenario – 50 MNOK/year and a high scenario – 70 MNOK/year.

The document is updated annually, and the previous document "Investments in the Norwegian eInfrastructure for Computational Science – An investment plan for the period 2007 – 2016", was presented in January 2007.

The document is written by the ReInfra committee with senior advisers Harald H. Simonsen and Gudmund Høst, The Research Council of Norway, acting as secretaries for the committee during the preparation of the report.

The ReInfra Committee has the following members:

Professor Knut Fægri, University of Oslo (Leader)
Professor Bjørn Hafskjold, NTNU
Professor Petter Bjørstad, University of Bergen
Professor Kenneth Ruud, University of Tromsø (to 31.12.2007)
IT-Director Roar Skålin, Met.no
Managing Director Kimmo Koski, CSC
Managing Director Petter Kongshaug, UNINETT (from 1.01.2008)

# Contents

# 1 Introduction

A modern eInfrastructure for computational science is an important part of the national research infrastructure. It is essential for doing high quality research in a large number of scientific fields. In particular, the national HPC infrastructure must provide the necessary resources and services to enable Norwegian scientists to do research of the highest international standard in close collaboration with their colleagues in other countries. The purpose of this document is to provide advise on investments in the Norwegian eInfrastructure for computational science over the next ten years.

Today the responsibility for the national HPC infrastructure for academic use rests with the eScience program (eVITA) in the Research Council of Norway (RCN). The operational responsibility is delegated to UNINETT Sigma AS through the NOTUR II project financed by the RCN and a consortium involving the four pre-2004 universities and the Meteorological Institute (met.no).

At present NOTUR II encompasses equipment at four different sites, i.e. at the four partner universities (see below for more details). Resources on these installations are distributed partly through national quotas, administered by the Allocation Committee, partly as local quotas, administered locally.

To provide advice on the development of the national HPC infrastructure the eVITA Program Committee (ePC) has established an Advisory Committee on Norwegian HPC Infrastructure (ReInfra). ePC has given ReInfra the task of developing a national plan for investments in eInfrastructure for computational science for the ten year period 2007 – 2016. In particular the terms of reference state (the full terms of reference may be found in Appendix A):

> *Write a draft proposal for a national investment plan for eInfrastructure for the next 10 years. This document should be specific enough to be used as an advice on investments due to a possible budget increase from 2007. The budgetary freedom in the NOTUR II long term budget should be incorporated in the document*

However, before turning to a plan for investments in e-Infrastructure, there are some overriding issues of more general nature should be considered, issues that effectively constitute boundary conditions on any plan for investment in a Norwegian e-Infrastructure. The most important of these is the *cost-sharing model*. Today the equipment as well as operating and support services are financed jointly by RCN (through the NOTUR project) and the installation site. This is necessary in order to optimize the volume of financing. The *de facto* consequence of this cost sharing has been a distribution of resources over the sites corresponding to the financial contribution. As discussed in previous evaluations and reports, this spreading of resources has both advantages and disadvantages; the main advantage being local funding and commitment; the main disadvantage is reduced flexibility in assigning national responsibilities to the different sites.

Linked to the problem of cost sharing, is the important issue of *task differentiation over sites*. This has been attempted in previous organization of HPC schemes in Norway, but mainly with a technology basis — different sites would take responsibility for MPP, clusters or vector technology respectively. With a more convergent technology picture, an alternative would be to differentiate over tasks. One may also envision certain types of application software being restricted to special platforms. Today, there is little or no such differentiation.

On a somewhat different level is the question of *the national roles played by the various participants in Norwegian HPC*. These include the RCN through the eVita program committee and its subcommittee ReInfra, UNINETT Sigma with the UNINETT Sigma Board and the UNINETT Sigma Advisory Committee. Other important role players are the NOTUR II Consortium partners and the program committees for those RCN programs that depend on HPC for the research funded by that program. The many different stakeholders in Norwegian HPC all have a say when the long-term strategy is worked out and implemented. Consequently, the decision-making process is lengthy. The hardware investments, which have to be made in order to keep the computational science community operational, then run the risk of being made on a short-term basis that is not necessarily well balanced. The present document attempts to contribute to a long-term strategy such that short-term decisions can be made in a way that is consistent with the long term strategy.

Related to this are a number of areas where responsibilities between the various actors are not clearly defined. An example of this is the cost of Norwegian computing activities arising more or less directly from the membership in CERN. Currently these activities are mostly funded by the RCN through special allocations to CERN related research, but no clear policy exists as to how much of this is a national responsibility and how much should be the responsibility of the community using these services. Similar situations exist also in other fields of research with unclear responsibilities regarding computational requirements arising from activities within large national or international collaborations and research programs.

While the three issues discussed above are mainly organizational, there is also an underlying issue of defining what *infrastructure* really means in this context, and where the border between infrastructure and science goes. Straight technology — computers, disk drives and other hardware — clearly classifies as infrastructure (at least in the context of Norwegian technical development in HPC, which is minimal). However, there are other areas where the distinction is less clear. An example is Grid development which may be viewed as a technical issue, but which also involves a large component of scientific development, both in the Grid middleware, but also for the user in adapting applications to Grid use. There are many such subareas within HPC infrastructure that carry a considerable component of science, not necessarily publishable, at least not in its own right. We feel that the funding of such tasks is important and must be taken care of within the eVita program. However, we also feel that infrastructure investments should apply to technology that is operational without further development. If infrastructure funding is used for development work, the development should have a clear objective to improve the user community's efficiency in computational science.

The lack of some important boundary conditions makes long term planning of Norwegian HPC investments a rather uncertain proposition. One problem is that political decisions need to be made in a number of areas. An example is the extent to which one may deviate from an equipartitioning of resources between the major sites of the system. This question needs to be considered in the light of both competence (what is the minimum amount of HPC required at a site to maintain competence?) and financing (can we expect an institution to contribute towards investment at another site?) To provide a basis for decisions on questions such as these requires discussions both with the site managers (IT directors) and the institutions.

The question of financing also points to another difficulty in Norwegian HPC planning. For nearly ten years the annual budget for high performance computing has been held nearly constant at approximately 22 mill. NOK. As a consequence the development of the HPC infrastructure has at times been driven more by opportunity rather than strategy. This is somewhat connected to the question of flexibility in procurements. In order to obtain an advantageous bargaining position, both with local institutions and with vendors, it is desirable to have rather large degree of freedom in procurement processes. However, this may lead to purchases that are cost efficient, but not necessarily strategically optimal. Thus, there is a need for making long term priorities and sticking to them as financing may have to be spread over more than one year. A further complication is that hardware being by far the most expensive items also has a dimensioning effect on the other activities.

Long term planning is not made easier by processes such as the sudden decision by the Ministry of Education and Research in early 2007 to reallocate 200 mill. NOK from the budget of the Research Council to investments in infrastructure for research. eInfrastructure was given 70 mill. NOK (out of the 200 mill) with conditions that the money had to be spent in the 2007 budget year. This actually meant that the total budget for eInfrastructure was just short of 100 mill. NOK in 2007. The extra budget resulted in large investments at all centers and nearly all the main systems were upgraded. Unfortunately, this seems to have been a one time increase, and in 2008 we are back to approximately 26 mill. NOK (including the Grid and storage activities). A major challenge will be to fund the operation of the systems at a satisfactory level and still be able to find the budget to fund necessary upgrades in the future.

The present document has a ten year time frame. However, it should be emphasized that ten years is a very long planning horizon for investments in technologies that are rapidly changing. Changing budgetary constraints further complicate the issue. For this reasons an investment plan can only give rough directions that will need to be carefully revised as the road ahead becomes clearer.

# 2 Analysis

In this chapter we present an analysis of the present situation for eInfrastructure for computational science in Norway. The analysis leads to suggestions as to where future activities within these areas should be directed. This forms the basis for the investment plan proposed in the next chapter.

Similar to the NOTUR II Long Term Plan[1] we define eInfrastructure in this context as:

- Hardware including operations
  - High end computational resources
  - Storage facilities
  - High-speed network
- Software including
  - System software and basic tools
  - Application software
  - Grid middleware
- Support
  - Basic help-desk support
  - Advanced user support
- Services for
  - End-user functionality
  - Performance guarantees
  - Quality assurance

In our discussion below we do not address services explicitly. However, we emphasize that the services provided are an important aspect a well functioning eInfrastructure for computational science. It is the responsibility of the NOTUR Metacenter to provide these services to the end users.

## 2.1 Important trends

### 2.1.1 Usage trends

#### *Scientific usage trends*

Computational modeling is today well established in many fields of research as an integral component of the research activities, and it is often referred to as the third way of scientific research, complementary to the traditional research methods of theory development and experiments. In addition to providing insight into scientific problems

[1] NOTUR II Long Term Plan 2006, Version 18.10.2006 by UNINETT Sigma AS.

only obtainable indirectly from experiments, computational modeling is in many cases the only possible approach for addressing important scientific questions. This is perhaps most clearly illustrated in climate research where predictions of the future climate on the basis of different political choices with respect the use of e.g. fossil fuels, is only possible through computational modeling. In a similar manner, the details of a chemical reaction is still outside experimental reach, but modeling can provide detailed insight into the mechanisms driving the reaction, and at the same time the accuracy of models can be benchmarked against available thermo-chemical and kinetic data.

The importance of computational modeling is well described in the recommendations of the President's Information Technology Advisory Committee (PITAC), who in 2005 noted that:

> *Computational science is now indispensable to the solution of complex problems in every sector, from traditional science and engineering domains to such key areas as national security, public health and economic innovation. Advances in computing and connectivity make it possible to develop computational models and capture and analyze unprecedented amounts of experimental and observational data to address problems previously deemed intractable or beyond imagination.*

Several of the large strategic research programs initiated by the Research Council of Norway are large users of computing powers. In particular, NORKLIMA, NANOMAT and FUGE are programs that are, and will remain, large consumers of supercomputing resources, and within all these research fields the need for computing resources is expected to increase in the future. RENERGI, PETROMAKS and CLIMIT may also be expected to be users of computing resources, though most likely to a lesser extent than the other three programs. The Norwegian eScience program (eVITA) recently initiated is not expected to directly be a large user of computing infrastructure. However, the program will hopefully lead to the development of improved computational algorithms and thus improved use of the advanced computational infrastructure. As a result of eVITA, computational science will advance, and the impact of computational resources will increase. The program may also need dedicated access to computers in the NOTUR program in order to explore the best computational algorithms for important scientific problems.

At least 3 of the established centers of excellence are also large users of national supercomputer resources, and several of the new centers include significant modeling activities. In general, an increased demand for computing resources driven by the various initiatives of the Research Council of Norway can be expected in the future.

The need for a European HPC infrastructure was recognized in the European Strategy Forum for Research Infrastructures (ESFRI) roadmap (2006). PRACE, the Partnership for Advanced Computing in Europe, intends to follow up the roadmap by creating a pan-European high performance computing (HPC) service consisting of major supercomputing centers. This infrastructure will be managed as a single European entity. European scientists and technologists will be provided world-class leadership supercomputers with capabilities equal to or better than those available in the USA and Japan. The service will comprise three to five superior HPC centers strengthened by regional and national supercomputing centers working in tight collaboration through grid technologies.

Another important development in Europe is the European Grid Initiative (EGI). The EGI Design Study represents an effort to establish a sustainable grid infrastructure in Europe. The main foundation of EGI is the National Grid Initiatives (NGI), which operate the grid infrastructures in each country.

### *Industrial usage trends*[2]

Computer simulation has rapidly become necessary for a large part of the industry. Simulation plays an important role in the design of materials, manufacturing processes, and products. Increasingly, computer simulation is replacing physical tests to ensure product reliability and quality. Fewer tests mean fewer prototypes, and the result is a shorter design cycle. Steady reductions in design cycles, in turn, are crucial to remain competetive in a world where the pace at which new consumer products are being developed is increasing every day. A modern eInfrastructure for computer simulation has become a necessary tool for an increasing part of our industry.

## 2.1.2 Technology trends[3]

This section provides a brief description of the most important technology trends that drives the development of the eInfrastructure for computational science.

### *Computer Architectures*

HPC architectures have evolved rapidly over the last 30 years. Technological developments combined with algorithmic and methodical developments have driven down the price/performance of HPC systems. The mainframes of the 70's and 80's gave way to the massively parallel systems and workstation 'farms' of the early nineties. These, in turn, have given way to the clusters of commodity nodes, most likely the primary form of HPC resources for at least the near future.

The focus on clusters and the quest to use them for tightly coupled applications have led to the use of high-performance interconnects. The difference in bandwidth between interconnects is decreasing, but in some cases the lower latency of specialized HPC interconnects is important.

[2] A similar argument is given in the report "Simulation –Based Engineering Science" by National Science Foundation's Blue Ribbon Panel on Simulation-Based Engineering Science.

[3] Some of this material is based on the report "The Swedish HPC Landscape 2006-2009 – Visions and Road Maps", April 2006 by the Swedish National Infrastructure for Computing.

Another clear trend is towards nodes containing an increasing number of processing units. Today, clusters built from nodes containing 4-16 processing units are commonplace. Such systems can provide improved performance per unit floor area and their nodes can be used in a flexible way. Each such node is typically equipped with more memory than the traditional 1 CPU node and is also capable of running shared-memory applications. This makes each node usable for single-threaded applications, with a high demand for memory, as well as for tightly coupled parallel applications with a modest scalability requirement.

Due to problems with cooling and energy consumption the clock frequency of processor chips seems to converge towards approximately 5 GHz using CMOS fabrication technology. This indicates that Moore's law will no longer be valid with regard to single processor speed. On the other hand, further advances in fabrication technology will make it possible to continue to increase the number of transistors on a chip according to Moore's law[4]. Accordingly, the current trend in processor architecture is to utilize the extra transistors for chip multithreading (CMT), where each processor is capable of executing multiple parallel threads. This can be achieved by introducing several cores per processor chip and/or by hardware support for several threads within a core. The technology has existed in high-end server processors for some time, but now commodity processors with multiple cores per chip and multiple threads per core are being introduced. Using the new processor chips in clusters will emphasize the trend that one node runs many threads. In order to accommodate the increasing number of threads, the memory capacity, memory bandwidth and the cluster network bandwidth per node will have to be increased significantly compared with today's nodes. *An important consequence of this trend is that one can no longer rely on increases in single processor speed alone to increase the execution speed of a program. An effort to parallelize the application is crucial to ensure a significant increase in execution speed.*

### Storage

Within the coming years we can assume a continuation of the development of increased capacity per disk. We can anticipate single disks with a storage capacity of several TB. The ongoing trend of investing in disks instead of tape will continue. However, some applications with very large and long-term storage demands will still require tape solutions.

In addition to the actual storage capacity, data transmission rates and I/O speed are crucial for efficient use of central storage facilities. High performance network technology and parallel file systems are ways to improve data transfer rates. Parallel file systems for cluster solutions can provide a very cost effective way to increase the bandwidth for data transfer from processors to disk. *It is evident that there is an urgent need for a national strategy and a uniform approach to the procurement and management of large scale storage.*

[4] This version of Moore's law states that the number of transistors on a chip approximately doubles every 18 months.

### *Networks*

Network bandwidth is increasing very quickly, and for most applications it is not a problem to transfer the data from site to site. However, for certain very data-intensive projects, the network technology will not provide sufficient data transfer capability. There is thus a splitting of applications into the ones for which data transport is not going to pose any problems and the ones for which data sizes are increasing faster than the available bandwidth.

To meet the need for both a high-capacity shared IP network and the need for point-to-point connections a new technology termed lambda networks or hybrid networks has evolved. This technique makes use of different wavelengths of light on optical fibers to create multiple channels on a single fiber pair; a technique known as Wave Division Multiplexing (WDM) or Dense Wave Division Multiplexing (DWDM) depending on the number of wavelengths (channels) supported. One channel (or wavelength) is used for interconnecting routers to create a general and shared IP network. The remaining channels are used to set up point-to-point connections as the need arises.

### *Grid Technologies*

Grid computing, storage and data management make it possible to bring together geographically dispersed resources and allocate them to specific applications. There is a wide range of Grid paradigms ranging from the loosely coupled Grids formed by individual PCs contributing unused cycles, to the fully planned and controlled Grids formed by connecting supercomputer centers. In these latter Grids the user has the advantage of a single sign-on to all resources and the redundancy of the grid system can be higher than if the user is limited to one or a few resources. Instead of accessing the resources by the traditional remote login procedure, the user interface is located at the desktop computer. The grid middleware provides authenticated communication between the services in the grid infrastructure.

There are a number of different middleware systems available, each providing a different grid flavor. In NDGF NorduGrid´s ARC and the European-developed LCG/EGEE middleware are used.

Grid technology and grid administration are still evolving rapidly and there are issues that have to be solved before the advantages afforded by a grid resource can be fully exploited. The development is, to a large extent, driven by research institutions or research projects and is supported politically as well as commercially. Major computer companies like IBM, Sun and Microsoft are pushing for grid standards. To reach the goal of a robust infrastructure it is necessary that a large number of institutions coordinate their efforts, and a further development towards middleware standards is expected in the next generation of grid solutions. This will certainly affect the development of both ARC and LCG.

## 2.2 Hardware

### 2.2.1 High-end computing resources

Large computing facilities are often divided into two groups: capability systems and capacity systems. *Capability systems* are computers that usually have many processors, a large (often shared) memory and low-latency, high-bandwidth interconnect between processors. These systems are typically used for time-critical or computationally very demanding applications that require the whole computer for a certain period. In Norway the weather forecast is run on such a system.

*Capacity systems* also offer large computational power but do not have the high performance interconnect that a capability system provides. This type of system is typically cluster based with a large number of nodes based on commodity PC or server microprocessors. They are used to provide a large job throughput and usually run a number of user applications simultaneously.

Due to a more expensive interconnect a capability system is usually more expensive than a capacity system with the same peak performance pr. processing element. This makes a capacity system an attracting alternative for the large number of HPC applications that do not need the high speed interconnect and large memory a capability system typically provides.

After the upgrades in 2007 the NOTUR Metacenter[5] consists as of April 2008 of the following systems:

| **Name** | **Available** | **System** | **Type** | **No of nodes** | **No of cores** | **CPU type** | **Peak (Tflops)** | **Total memory (GB)** | **Disk capacity (TB)** |
|---|---|---|---|---|---|---|---|---|---|
| hexagon | 02/2008 | Cray XT4 | MPP | 1388 | 5552 | Opteron | 51 | 6064 | 288 |
| njord | 12/2006 | IBM p575+ | Distri-buted SMP | 65 | 944 | Power5+ | 7,2 | 2272 | 120 |
| stallo | 01/2008 | HP BL 460c | cluster | 704 | 5632 | Xeon2 | 60 | 12064 | 128 |
| titan | 10/2007 | Sun X2200 | cluster | 307 | 2480 | Opteron | 21,5 | 5376 | 10 |

The systems are physically distributed in four national centers located in Oslo, Bergen, Trondheim and Tromsø. The computational resources can be accessed through the national academic network from any university. Due to the upgrades done in 2007 all of the systems can support HPC tasks by today's standards. This has not always been the case in the Norwegian HPC system.

[5] See http://www.notur.no for more information.

The computational needs of different research groups differ. Some tasks can be done on a local desktop/workstation or cluster while other tasks need a larger system. A balanced computing ecosystem requires computing resources at all levels from the local workstation to the national and international top-class performance systems. In addition to scalable hardware, focus on scalable software development, code optimization, efficient data management, high-speed networks and expertise, such as technical competence and expertise for scientific computing, are all needed to make this a well functioning ecosystem.

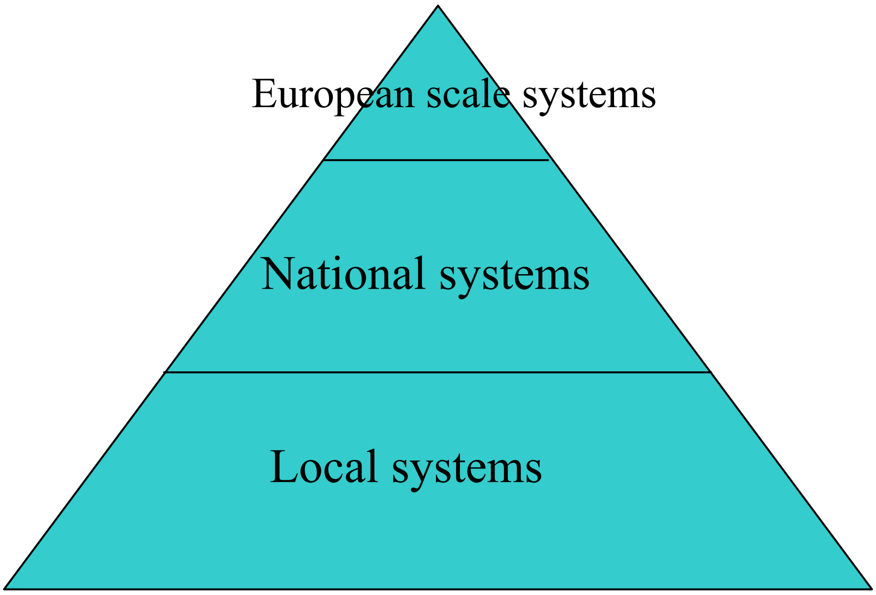

**Figure 1:** The Performance Pyramid

A balanced ecosystem can be described using a computational pyramid model. The research groups should be able to access resources at different levels depending on their needs and the quality of research. Not all the resources can or even need to be provided nationally, but it is important to have an option to access the top if needed. This requires international collaboration in addition to investments in local resources. It is very important to notice that without a solid and strong national HPC ecosystem with a sufficient computing infrastructure, it is not possible to utilize the peak of the pyramid resources efficiently, if at all.

To provide computational science in Norway with a balanced ecosystem giving access to resources from basic infrastructure to the systems at the top of the pyramid, it is necessary to develop a national strategy of computing which incorporates a better division of labor between the different national centers. This should include less overlap in computing services and an increased international collaboration to provide access to resources not currently available in Norway.

There are some important factors that influence investments in computing resources on the national level. These are:

1. The lifetime of a new computer system.
   Due to Moore's law a lifetime of more than 4 years not advisable.
2. The needs of met.no for operational forecasting.
   Today these needs imply that there must be a system of a minimum size in the NOTUR Metacenter.
3. The advantage of spreading investments in computer hardware over time.
   This is also a consequence of Moore's law. If all investments are done at the same time, all systems will be outdated at the same time. Spreading investments over time (for example 2 years) will guarantee that there will always be a national system that is reasonably up to date.
4. Not all use requires a system with a high speed interconnect (capability system). Using clusters, that are often more cost effective (capacity systems), for such use may free funds for investments in other parts of the national eInfrastructure .

These factors have some implications with regard to investments on the national level given different budget scenarios.

*Low budget scenario (22 MNOK/year through eVITA)*
Unfortunately, after the investments made in 2007 this is more or less the situation in 2008. Priority should be given to operate the systems introduced.

There should be no need for further investments before late 2009. Since njord, which is the system used by met.no for forecasting, is the oldest system, the next investment should then be in a new system that can replace njord and handle the needs of met.no.

An upgrade of one of the other three systems (not used for forecasting by met.no) should be considered for late 2010.

A new system should then be introduced in 2012 and so on. It is worth noting that if a distributed model for a capacity system is chosen, one has the possibility of spacing the investments in the partial systems that make up the capacity system over time. This will ensure that at least some part of the system is reasonably up to date.

Access to top-end resources in the pyramid will be difficult to fund. In particular, funding for access to high-end systems outside Norway would have to come from the projects that need such access.

*High budget scenario (70 MNOK/year through eVITA)*
This budget scenario will make it possible to consider three alternative investment strategies.

1. Use the same strategy as for the low budget scenarios, but increase the size of the systems.

2. Invest in a high end capability system that is upgraded after 1 ½ – 2 years. Invest in a new system to handle capacity use every two years. There will be three national computer resources available. Both the capability system and at least one of the capacity systems will be reasonably up to date.
3. Invest in a new capability and capacity system every two years. There will be four national computer resources available where at least two are reasonably up to date.

Of these alternative investment strategies, number 2 has the advantages that:

- It provides a reasonably up to date capability system.
- The size of the system would make it more attractive for vendors which hopefully will make it possible to negotiate a better prize.
- At least one of the capacity systems will be reasonably up to date.

Compared to alternative 3, alternative 2 has a slight disadvantage in the sense that there will only be a single capability system. Downtime on this system will be more critical than downtime on one of the two systems in alternative 3. Compared to alternative 1 the disadvantage is that the largest capability system will be of lesser power for the first 1 ½ years. However, this will to some extent be compensated by the time a user needs to adapt to a new system.

We note that alternative 3 is closest to the present situation.

Access to high-end computing resources outside Norway should be negotiated for those research groups that need such access. In particular, Norway should ensure that Norwegian researchers will have access to European level computing resources if/when they get established.

### 2.2.2 Storage facilities

The need for large data storage has increased considerably over the last years, especially in areas like geosciences (climate modeling) and physics (high energy physics). For certain areas, high-resolution data is collected from real-time instruments (e.g., sensors) and large complex distributed databases are used. In other cases, large quantities of data are being generated during long computer simulations. For many of these cases, data cannot easily be regenerated and must be stored (archived) over longer periods of time.

One must distinguish between different types of storage. Factors that must be taken into account are the data set sizes and their composition (e.g., granularity), the complexity of the data sets (e.g., flat UNIX-like file systems or complex hierarchical databases), the value of the data (e.g., redundancy and backups), the validity/expiration of the data (e.g., duration of the storage), and data access patterns (e.g., access frequency or need to access subsets of the data). Three major classes can be identified:

Class 1 **Temporary Storage:** Fast direct attached storage. The data is stored only for a short period of time, typically the duration of the simulation. The performance is in balance with the computing resource, and the granularity of access in on block level (common file I/O access).

Class 2 **Permanent storage:** Direct attached, SAN or NAS storage. The data is used actively in one or more projects and cannot be easily regenerated. The storage may be accessible from every computing resource that is used in the project, and the granularity of transfers is on file level.

Class 3 **Long-term storage:** Large and stable data repository, typically slow disk or tape. The data is not used actively in the project, but may be accessed. This can also be data stored for legal reasons. The data may be accessible from a file system or an (web-based) interface, and the granularity of transfers in on the size of datasets.

Long-term storage must be guaranteed for more than 10 years and must be able to cope with shifts in storage technology.

Storage capacity within the NOTUR Metacenter is based on computational resources with local devices for temporary and permanent storage and for backup of data. In total, the NOTUR Metacenter has in excess of 600 TB of disk for temporary and permanent storage. In addition, a storage system has been established for the Nordic Data Grid Facility (NDGF). The Norwegian storage contribution to the center is approximately 350 TB of disk storage (March 2008). The need for large amounts of storage today appears to be concentrated to a few areas, in particular geophysics and high energy physics. In the future other areas (e.g. medical equipment) may also generate large needs for storage.

The NorStore[6] project has recently been established by the eVITA program to handle the needs for long term data storage and data management from several fields within the natural sciences. The project envisages the creation of a permanent national infrastructure that will consist of several large data repositories with a core set of services, standards and interfaces that will be maintained across the infrastructure. The initial infrastructure will consist of two facilities, each of about 600 TB net capacity based on StorageTek 6540 technology.

In a *low budget scenario* priority should be given to operate the existing facilities in NorStore. Investments in new repositories should only be done if they can be funded by the research projects that need the repositories.

In a *high budget scenario* we also recommend investment in new repositories.

[6] See http://www.norstore.no for more information.

### 2.2.3 High-speed network

A reliable high-speed network is essential for the use and operation of the Norwegian eInfrastructure for computational science. Such a network is also essential for participation in and interaction with the international research community.

UNINETT AS is the main operator of the academic network in Norway. Today, the links between the four university partners operates at 2,5 GBit/s. Hopefully, these links will be upgraded to 10 GBit/s during 2008.

Since the responsibility for the development of this network rests with UNINETT AS, we will not give specific recommendation regarding investments in the network in this document. We do stress that the planned upgrade in capacity is timely and essential for the continued development of the other parts of the Norwegian eInfrastructure.

## 2.3 Software

### 2.3.1 System software and basic tools

This software includes basic system software for operation and resource monitoring, and basic tools to enable the execution of applications in a distributed infrastructure. We also include end-user tools for the development and analysis of applications.

Typically, the basic system software is Unix-based for capability systems with Linux becoming more and more popular for the cluster-based capacity systems. Even though it would be tempting to try to standardize on Linux, it will probably not be wise to do so for performance reasons.

With regard to basic tools the situation is different and it would be preferable to standardize as far as possible the tools that are used in the Metacenter. Some standardization exists, but we think this can be done more extensively.

### 2.3.2 Application software

Scientific applications are invariably coupled to software in the form of one or more computer programs. These come in three main varieties:

- Homemade, user developed programs
- Public domain software (freeware, shareware) distributed by groups at other institutions or by international collaborations
- Commercial software produced and sold by vendors who may either be straight business ventures or a research group (or groups) trying to generate income from their science.

Only software in this last category will appear directly in the budgeting process. Up to now the NOTUR system has managed to cater to the various users' needs within the normal budgets, and this has not been a major expense item. In the short run we foresee no dramatic changes beyond a general price increase. It is therefore necessary to continuously monitor license expenses and negotiate for advantageous price agreements. This is difficult, as commercial application software often occupies a semi-monopoly position.

In the somewhat longer run, there may be a shift towards more commercial software. This is a trend that has been observed over the years in Chemistry, and it would not be unreasonable to expect something similar for areas like Medicine and Biology as they mature. This could lead to a need for more stringent routines with regard to choice of applications to support. It would also place a further importance on license handling procedures.

## 2.4 Grid

Grid technology may be seen as an effort to move the various user interfaces to computing resources (as well as other large scale scientific instruments) away from the resource itself. If successful, this may result in higher scientific productivity. The scientists can do computing as well as management of data (possibly from a large number of sources) in the most efficient manner via a standard desktop interface regardless of the specific task or computing resource. There are, however, a large number of issues that need to be addressed before grid technology can deliver this vision. These issues are technical, administrative, political as well as the training people to change their way of working. This last point is crucial, but advances here require that scientists are motivated to change their way of work. This, in turn, is hard to accomplish unless there are pretty obvious and direct benefits to the individual.

The NorGrid[7] project is established to provide grid-based services and to "grid-enable" scientific applications on top of the infrastructure for computational science in Norway. The long term vision of the project is that many compute and storage resources will be accessed through a grid interface. The services and interfaces should provide easy and secure access to distributed resources and help researchers create and participate in computational challenges of scope and size unreachable on single facilities alone. NorGrid is currently funded with 2,5 MNOK from the Research Council for the period 2008-2010.

For the planning period, it is important that Norway develops a basic Grid infrastructure in order to follow the international and Nordic developments. In particular, the activity related to the NDGF role as a Nordic Tier-1 Grid Center, must be followed up with adequate funding on a national level. In order to do so the NorGrid project should be given the resources to establish the basic services needed to run the grid. The main computing and storage resources in NorGrid should be the same as in the NOTUR Metacenter.

[7] See http://www.norgrid.no for more information.

## 2.5 User support

User support is a very labour-intensive activity with the overall objective to assist users with various technical problems and improve the effective utilization of the hardware resources. Typically, the annual user support budget in NOTUR has been approximately 3,6 MNOK, much of which is as in-kind contributions from the sites. A national helpdesk is used by the NOTUR Metacenter where users can submit queries and problems. The Metacenter staff provides assistance for a variety of issues, including the installation and compilation of software, the execution of applications, and any machine-specific issues. A website exists where users can track the history and status of their requests.

User support is here divided into three categories; (i) Help to get started, (ii) Application support and (iii) Data support.

### 2.5.1 Getting started and day-to-day support

The last NOTUR user survey shows a high satisfaction with the quality of the support given by the technical staff and the response times and follow-up to requests and problems. A moderate improvement may be made by better coordination between the sites in order to level the load on the support personnel. The quality of the NOTUR support web pages and other end-user documentation should be improved.

The NOTUR user is offered introductory courses and tutorials for new users and more specialized and/or discipline-specific courses for experienced users. Such courses should be continued and repeated regularly, especially targeting new users. The partners are encouraged to jointly develop and share course material.

### 2.5.2 Application support

Software lifetime is much longer than hardware lifetime. This means that a lot of effort is spent on porting codes, which may have been developed by several people over several years, to new platforms every few years. As a result, several software packages which have a long history of upgrades and modifications, have become inherently complex and hard to maintain, and may not use the nowadays accepted programming standards. In practice, such codes are hard to port to new platforms. The NOTUR project should provide support for users of such applications and consider the benefits and costs of maintaining the current software versus trying to improve it.

For a user to be in the front of computational science, the competitive edge may be improved by better adaptation of the software to the hardware and better algorithms. However, few users have the skill (or time) to adapt and optimize their codes themselves without help from advanced tools and people that are skilled in using them, which should be provided by NOTUR.

Advanced user support primarily targets end-user application software. Examples include complex application enabling, parallelization of software, performance tuning, and software development (e.g., new or improved functionalities and interfaces, adding or replacing software modules, etc.). An important element of advanced user support is to make a code run efficiently on a low-cost platform. NOTUR should continue the advanced support activity in close cooperation with leading user groups and with strategic applications areas. The interface between advanced user support and eVITA computational science development is as yet undefined, and should be clarified, not least for budgetary reasons. Research itself is not financed through the NOTUR project, but active cooperation between research and development should be established by shared financing of personnel or equipment.

### 2.5.3 Data support

The users should be offered advice on how to structure their data, whether they are generated by the users themselves of acquired from some other source. The professional service should include structuring data for effective and safe storage, retrieval, maintenance/restoration, and porting to new media.

Data storage may be spread over several units, making maintenance and retrieval difficult. Good tools for distributed data storage and retrieval are becoming available, and NOTUR should develop services for effective use of such facilities and tools.

In order to increase the value of stored data in trans-disciplinary fields, standards should be made in order to facilitate communication between bases and data retrieval from different bases.

Most users will not have the skills or system to generate quality metadata. Much user time can be saved with a better support system for data management, and NOTUR should develop this.

NOTUR should develop a strategy for data safety and communicate this to the users.

# 3. Recommendations and investment plan.

The terms of reference state that: "*ReInfra is asked to base their advice on the three budget scenarios given in the eVITA Program plan (with the actual 2006 budget numbers included).*" The-VITA program plan lists the following budget scenarios:

| | 2006 | 2007 | 2008 | 2009 | 2010 - 15 | Sum |
|---|---|---|---|---|---|---|
| Low budget | | | | | | |
| *Research* | 16 | 30 | 30 | 30 | 240 | 346 |
| *Infrastructure.* | 22 | 22 | 22 | 22 | 132 | 220 |
| Moderate budget | | | | | | |
| *Research* | 30 | 40 | 60 | 60 | 360 | 550 |
| *Infrastructure* | 35 | 50 | 50 | 50 | 210 | 395 |
| Recommended budget | | | | | | |
| *Research* | 50 | 80 | 100 | 100 | 600 | 930 |
| *Infrastructure* | 50 | 70 | 70 | 70 | 360 | 620 |

*Low, moderate and recommended budget frameworks (all figures in million NOK).*

Since then the budget numbers for 2007 have become available, and for infrastructure these are higher than the *high* alternative due to the extraoridinary budget increase of 70 mill. NOK for investments in eInfrastructure. However in 2008 the budget are back to approximately the *Low* alternative. There are some political indications that this might change to a higher level for coming years.

## 3.1 Low budget scenario (22 MNOK/year)

This scenario, which reflects the funding level today, gives limited opportunities for the development of a robust eInfrastructure for computational science.

**Hardware**

*High-end computing resources*

Priority must be given to operating the new systems introduced in 2007. A new system to handle the needs of met.no for operational forecasting should be introduced late 2009. An upgrade of one of the other systems should be done late 2010. Further ahead, new systems should preferably be introduced with a spacing of two years. The capacity system can be distributed[8], which would make it possible to space introduction of the partial systems. Funding for access to high-end resources outside Norway will have to come from the projects that need such access.

*Storage facilities*

Priority should be given to investment in temporary storage and permanent storage (SAN or NAS) that complement the computational resources. Investments in storage repositories outside what is already done in NorStore as a part of NDGF for CERN and NoSerC for climate research will be difficult without extra funding. This funding should come from the research projects that need the repositories.

### Software

*System software and basic tools*

We think it would be preferable to standardize as far as possible the tools that are used in the Metacenter. Some standardization exists, but we think this can be done more extensively

*Application software*

It is important that computational resources on the national level provide the necessary application software for scientific use. This has not been a significant part of the NOTUR budget so far and we recommend that this is kept on same level as today. Introduction of specific application portals should not be prioritized in this budget scenario.

### Grid

For the planning period, it is important that Norway develops a basic Grid infrastructure in order to follow the international and Nordic developments. In particular, the activity related to the NDGF role as a Nordic Tier-1 Grid Center, must be followed up with adequate funding on a national level. In order to do so NorGrid should be given resources to establish the necessary basic services needed to run the grid. The main computing and storage resources in NorGrid should be the same as in the NOTUR Metacenter. Development activities should not be prioritized under this budget scenario.

### User support

We see no possibility to increase the investment in user support under this budget scenario. In particular, it would be difficult to increase the level of advanced user support. In order to make users less dependant on the user support organization, priority should be given to improve web-based support.

[8] The degree of distribution of the capacity system will be further investigated in more detail by ReInfra and a recommendation will be made at a later date. This also applies to the medium budget scenario.

## 3.2 High budget scenario (70 MNOK/year)

This scenario, which reflects a reasonable budget scenario, should make it possible to develop a robust eInfrastructure for computational science.

**Hardware**

*High-end computing resources*

Priority must still be given to ensure that there are some relevant computing resources on the national level. We recommend an investment strategy based on alternative 2 in Section 2.2.1 where a capability system is introduced an upgraded after approximately 1 ½ year. A new capacity system should be introduced every 2nd year. With a four year life time for each system this implies three systems on the national level.

Access to high-end computing resources outside Norway should be negotiated for those research groups that need such access. In particular, Norway should ensure that Norwegian researchers will have access to European level computing resources if/when they become available.

*Storage facilities*

Like in the low budget scenario priority should be given to investment in temporary storage and permanent storage (SAN or NAS) that complement the computational resources. We also suggest the introduction of new data repositories though the NorStore project.

**Software**

In addition to what is proposed in the low budget scenario introduction of specific application portals would be possible in this budget scenario.

**Grid middleware**

In addition to what is proposed in the low budget scenario development of application/grid portals should be possible under this budget scenario.

**User support**

Like in the low budget scenario the Metacenter should be supported on at least the same level as today. The web-based support should be improved considerably. We would also recommend increased emphasis on advanced user support. In particular, support directed towards specific scientific fields. Support for parallelization/optimization of user applications should also be increased.

# Appendix A. Terms of reference

### eVITA – Specification of mandate for the Advisory Committee for Investment in eInfrastructure

In the mandate for the Advisory Committee for Investment in eInfrastructure (ReInfra) the main task for the committee is to advise the eVITA Program Committee (PC) on:

- Long term investment strategies for eInfrastructure (high performance computing, storage and GRIDs) in Norway
- International trends in technology
- The need for computing resources and advanced user support for all Norwegian academic research and operational weather forecasting

ReInfra shall also advise on the need for eInfrastructure to support prioritised research areas including the strategic programmes in the Research Council.

ReInfra shall base their advice on the existing national eInfrastructure (high performance computing, storage and GRIDs) and how this infrastructure is organized. The advice shall incorporate the needs of existing user groups and facilitate use of the resources by new users from other scientific fields.

More specifically, the eVITA PC would like ReInfra to advise on the use of a possible increase in the budget for eInfrastructure investments from 2007.

ReInfra should base their advice on the premise that NOTUR II have had their plans for 2006 accepted including plans for a new machine in 2006/2007. Possible changes in the plans for investments through NOTUR II are not  realistic until after 2007.

In addition to what is stated above, the eVITA PC would like ReInfra to specifically address the following:

- What is the potential for regional (Nordic) and international cooperation when it comes to investments and user support?
- What kind of eInfrastructure should be prioritised in order to best serve the needs of the Norwegian research community and the operational forecasting at met.no?
- What should be the balance between investments in infrastructure (HW/SW) and user support (including advanced user support)?

In addressing the questions stated above the following should be addressed

- International development of eInfrastructure, particularly in Europe

- The development of pan-Nordic eInfrastructure, like the Nordic Data Grid Facility
- Technological trends
- Specific national user needs
- Institutional needs and investments strategies

ReInfra is asked to base their advice on the three budget scenarios given in the eVITA Program plan (with the actual 2006 budget numbers included).

The eVITA PC would like the ReInfra to adhere to the following milestones.

1. Assess the NOTUR II investment plan with focus on the consequences given by investments in 2007. **Due June 5th, 2006.**
2. Give an oral presentation of alternative directions for investments in eInfrastructures for the next 10 year. To be given in person on the PC meeting June 19th, 2006. The purpose of the presentation is to initiate the strategy discussion in the PC.
3. Write a "List of Opportunities" document, identifying a set of optimizing conditions and critical investment opportunities in the eInfrastructures area for the next 10 years. The document should give a high level overview of the requirements for a cost-efficient and healthy national eInfrastructure, suggesting opportunities for science with regard to questions stated above. **Due November 6th, 2006**.
4. Write a draft proposal for a national investment plan for eInfrastructure for the next 10 years. This document should be specific enough to be used as an advice on investments due to a possible budget increase from 2007. The budgetary freedom in the NOTUR II long term budget should be incorporated in the document. **Due November 6th, 2006**.

The eVITA PC suggests the following yearly cycle for the work in ReInfra, **starting from 2007:**

1. Write an eInfrastructure Roadmap for Norway for the next 10 years. To be revised annually, and/or extended by in-depth analyses on specific topics. **Due May/June**.
2. Assess the current NOTUR II investment plan with focus on the consequences given by investments in the following year. **Due May/June.**
3. Propose a long term national investment plan for eInfrastructure for the next 10 years. To be revised annually, based on current budget estimates. The document will be input to eVITA PC budgetary discussions. The budgetary freedom in the NOTUR II long term budget should be incorporated in the document. **Due October/November**.

# Appendix B. The Performance pyramid.

In this appendix we present a more detailed description of the Performance pyramid for scientific computing. The figure, presented above in section 3.1.1 is repeated here for convenience.

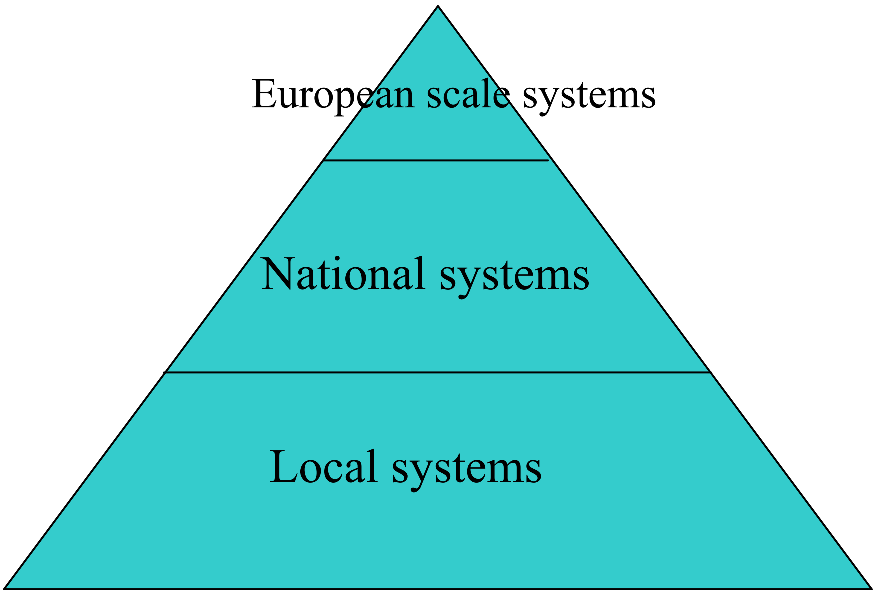

**Figure 1:** The Performance Pyramid

**Levels of the computational pyramid**

*Local systems*

The amount of local resources in terms of computing hardware and related technical expertise is due to the distributed national centers, one of the strengths in the Norwegian system. Providing NOTUR funding to universities attract local funding in addition and enables synergy in maintaining local resources. In addition to the hardware, technical competence is distributed widely through multiple centers.

It is important to keep up with the development and continue to support local computing resources on a suitable level. In addition, it is important to continue training new computational experts from the universities through visible presence of national HPC centers in four university locations. Attracting more young people to jobs within computational science domain promotes renewal and new ideas.

*National systems*
National scale computing resources are available in four locations: Oslo, Bergen, Trondheim and Tromsø. The national funding for these systems is increased by local university and research center funding, which at the same time increases synergy and cost-efficiency for operations. However, due to a distribution of resources it is challenging to reach performance levels high enough in the computational pyramid, and the danger to repeat the same level of moderate scale performance in all locations exist. More effective profiling of the national centers through available hardware and software environments and expertise in computational science is needed. Applications require different types of systems depending on the algorithms used and implementation, such as massively parallel systems with specific low-latency interconnects or standard network interfaces, systems with a large shared memory or fast storage systems. Optimal workload division between the centers allows cost efficient operations and new opportunities such as optimization in software licensing usage/cost and time sharing of limited technical and scientific competencies.

*European scale systemss*
Currently, there are not many Norwegian user groups who require extreme computer power in order to do advances in their scientific field. A few advanced researchers with a strong HPC background do exist, but an investment of 10-100 MEUR for hardware systems would probably not be justified only for a small number of people. However, it is important to facilitate access to appropriate resources for high-quality scientific projects if that is needed.

More active participation in international projects will allow better possibilities for access to top-class computing resources. In addition, opportunities for direct collaboration and resource sharing within the Nordic region or with other European centers in HPC utilization should be promoted.

*Interactions between layers*
In accessing European level resources it is important to notice that efficient utilization of top-class supercomputers require a strong national infrastructure to support the actions. Specifically, moderate scale systems are required  to run medium scale computations, for testing and building up the models which could then be scaled to 1000 or more processors. The hardware systems of various levels in the performance pyramid do not solve the problems without strong enablers of the performance, such as scalable software development and competent people.

The borders between different layers of performance can not be defined accurately. This is not necessarily a problem, but it is needed to focus on how the user can utilize services of different levels as transparently as possible. The smooth utilization is a challenge which is not purely technical – issues such as contracts, software licenses, authentications and peer-review processes for deciding on utilization of limited resources play a key role.

Practical and technical limitations might also be set by data storage. Even if accessing HPC systems remotely would be easy, transferring the required data across the network can be challenging. For data intensive computing the workload division might require special actions, such as investments in networking technology.

ISBN 978-82-12-02594-3 (printed version)
ISBN 978-82-12-02595-0 (pdf)