The Allocation of HPC-resources

Processing and distributing language data

Three-dimensional numerical modeling of water and sediment flow in rivers

The Norwegian metacenter for computational science
CONTENTS

The dynamics of ocean currents:
An interaction between processes from small to large length scales p : 4

Computer simulation of genome expression and regulation
– from yeast cell division to human cancer p : 7

Alloy design p : 10

Processing and distributing language data p : 12

Three-dimensional numerical modelling of water and sediment flow in rivers p : 15

The Resource Allocation process p : 18

Cover picture: Photograph of the delta of Lake Øyern, Norway.
Copyright: Fjellanger Widerøe AS

Number 2 : 2008
A magazine published by the Notur II project – The Norwegian metacenter for computational science

Editorial Director: Jacko Koster
Contributing Editor: Eva Irene Haugen
Subscription: An electronic version is available on www.notur.no. Here you can download a PDF-file of the magazine and subscribe or unsubscribe to the magazine.
ISSN: 190-1956

Reproduction in whole or in part without written permission is strictly prohibited.
e-mail address: sigma@uninett.no. Phone: +47 73 55 79 00. Fax: +47 73 55 79 01.
Print: Skipnes

Disclaimer: UNINETT Sigma is not responsible for errors, opinions or views expressed by the authors or the contents of external information referred to in this magazine.
EDITORIAL

Assuming there have not been last-minute changes to the plans of CERN as they were known at the time of writing this piece, the first beam of particles has been circulated in the Large Hadron Collider (LHC) on September 10. Once stable circulating beams have been established, they will be brought into collision, and the next step will be to boost the energy in these collisions. The LHC is the world’s most powerful particle accelerator and one of the largest scientific undertakings ever. Once LHC reaches design performance, probably by 2010, it will produce beams around 30 times more intense than any previous machine. The energies in the collisions may go up to around 1000 TeV (Tera electron Volt) depending on what matter is smashed into each other.

Energy levels of the order of 1000 TeV sound huge, but in absolute terms, these energies are not impressive. In fact, as CERN pointed out on its web pages, 1 TeV is about the energy of motion of a flying mosquito. However, the concentration of the energy is what makes LHC so interesting. By colliding particles, LHC squeezes many TeVs into an area about a million million times smaller than a mosquito. This creates conditions where physics goes bananas, particles will break up into their smallest fragments, thereby hopefully revealing properties that many physicists have assumed to exist but never observed in practice and/or revealing new and/or unexpected phenomena. In all cases, LHC will take particle physics research to a new frontier.

The LHC is expected to produce around 15 Petabytes of data each year for the next 10 to 15 years. This data will be distributed, managed and analysed via a world-wide WLCG computing grid. The Nordic countries have joined forces and created the Nordic Data Grid Facility that coordinates the operational tasks for the infrastructure resources that are set up by the individual countries. Alone in Norway, around 1.5 Petabyte LHC data will be stored by 2010 at the University of Bergen and the University of Oslo and about a thousand processor cores will be dedicated to the analysis of LHC data (although in practice the setups allow that many more processor cores can be involved). These numbers will increase significantly in the years after.

The start-up of LHC represents a major milestone for the whole European grid community, as LHC has been the driving force behind grid infrastructure. Large European initiatives like Enabling Grids for E-sciencE (EGEE), Nordic collaborations like the Nordic Data Grid Facility and the Norwegian grid initiative NorGrid would probably not have been created if LHC would not exist. Hundreds of scientists have contributed in the past ten years to make the WLCG grid ready for its purpose. In the past couple of years, preliminary versions of the grid infrastructure have been going through periods with increasingly intensive testing, but soon the first real production phase will start.

After LHC and WLCG have entered ‘normal’ production in some years from today, high-energy physics will continue to play a significant role in further grid innovation. There will still be a considerable need for further development, including the rethinking of components of the infrastructure that had to be delivered with what turned out to be a challenging time schedule. In addition, grid technologies and infrastructures in many different forms and implementations have also been adopted by other scientific communities with an international researcher base. It will be a challenge to make these infrastructures interoperable such that resources and knowledge can increasingly be shared between these communities.

Jacko Koster, Project Coordinator Notur II, Managing Director UNINETT Sigma AS

FEATURES

The study of genome expression and regulation via computer simulation may change the way molecular biologists do research in the future.

Norway produces over one million tons of aluminum alloys per year, and the industry is eager to understand more of the fundamental physics in this field.

The advancement of fast computational resources has in recent years made it possible to calculate water flow in rivers in three dimensions.
The dynamics of ocean currents:
An interaction between processes from small to large length scales

The dynamics controlling ocean currents are not fully understood. The large-scale circulation depends on smaller scale processes that are not resolved in global or large-scale ocean circulation models. An understanding of these smaller scale processes and their effect on the large-scale circulation is highly important for improving the performance of ocean circulation models.

Modeling ocean currents
Limited computer resources put major restrictions on modeling the world ocean circulation. This forces modelers to use coarse model grids which do not resolve processes of high importance for the dynamics of the circulation. Large-scale ocean currents (like the Gulf Stream or the Antarctic Circumpolar Current) are characterized by horizontal length scales of several thousands of kilometers. However, their dynamics are to a large degree determined by processes characterized by horizontal length scales of the order of 10 km and even smaller. A realistic model of the ocean currents therefore requires huge computer resources as one needs to resolve horizontal length scales of 10 km while the size of the model domain is several thousands of kilometers across. Other requirements, such as resolution in the vertical coordinate and time come in addition. The result is that for global or large-scale ocean models, it is impossible to resolve the processes controlling the ocean currents in a straightforward manner. To get around this problem, ocean models involve parameterizations of processes with length scales similar or smaller than the grid size, such as mesoscale eddy activity (~10km) and turbulence (scales from millimeters to tens of meters). However, many of the smaller scale processes are still poorly understood, and parameterizations of these processes in ocean models may therefore not be a good representation of the actual physical processes. The parameterizations may be crucial for the model results, and in order to improve them a better understanding of the processes is required.

At the Norwegian Polar Institute (NPI), we work on gaining insight into the physics controlling the dynamics of the ocean circulation. To describe our work, we first introduce some basic dynamics of ocean circulation.

Large-scale ocean dynamics and rotation of the Earth
The rotation of the Earth is a very important parameter for the dynamics of ocean circulation. To understand this, we start by taking a look at a spinning top. Whether a spinning top will stand upright or not depends on its angular momentum. The angular momentum of a spinning top is given by its rotation speed and width. A wide spinning top has larger angular momentum than a thin spinning top, if the rotation speed is similar for the two. Therefore, it is easier to get a wide spinning top to stand upright than it is with a thin. This is the reason why spinning tops are constructed like they are: thin at the top (this makes it easier to spin it up fast) and wide below (this gives it a large angular momentum). The importance of this shape can be seen by trying to make a pencil stand upright by spinning it up. It is practically impossible to manually spin it up fast enough.

The link between a spinning top and the ocean dynamics is seen when considering the effect of the rotation of the Earth. This effect can be observed in a rotating tank. Consider a tank filled with water that is set in rotation around a vertical axis. After a while, all the water will rotate with the tank. If we do not disturb the water in the tank in any way, the motion of the water relative to the tank will be zero. For an observer...
The dynamics of ocean currents:
An interaction between processes from small to large length scales

rotating with the tank, the water looks like it has no motion. However, in reality the water is moving with the same rotation speed as the tank. Any vertical column of water in the tank will rotate similarly. When the water in the tank is disturbed and set in motion relative to the tank, the rotation has a major influence on the circulation. Due to the rotation, vertical columns of water will stand upright in a similar way as a spinning top. The circulation in the tank will then be as illustrated by Figure 1. Since the water columns are not tilting, vertical motions will be suppressed and the circulation will be two-dimensional in the horizontal plane with no variations in the vertical.

However, the circulation in the tank consists of circulation features with different length scales. This is clearly illustrated by Figure 2a, which shows a snapshot of surface temperature in a simulation of the circulation in a circular basin. We see that the surface temperature consists of features of both small and large scales. The small scale features, called eddies, are constantly varying in time. Their spatial scale is about 20 km. On a larger scale, the temperature is warm near the boundary and cold in the center of the basin. This is a large-scale feature with length scale similar to the radius of the basin (200 km). The circulation in the basin contains the same length scales as the temperature. A large-scale circulation moving counterclockwise around the basin in addition to strong circulations associated with the smaller eddies.

The difference between the dynamics of the mesoscale eddies and the large scale circulation can be understood by comparing a thin and a wide spinning top. The eddies are like thin spinning tops. Inside the eddies, vertical columns of water may tilt and vertical velocities are allowed. The large-scale circulation behaves like a wide spinning top. On a large scale, water columns will stand upright, as illustrated by Figure 1. This large-scale motion is therefore strictly horizontal.

The main point to remember from this section is that vertical motion occurs on length scales of the order of 10 km, while large-scale (>100 km) motion is strictly horizontal. In the following, we describe our results and projects where mesoscale motion is essential for the dynamics of the large-scale circulation.

Topographic steering
An important effect of the suppression of vertical velocities due to rotation is that large-scale flow must follow a path of constant water depth. The reason for this is that a vertical column of water moving from shallow to deep water (or opposite) will be stretched (or compressed), which requires vertical velocities. This effect has been known for a long time. A hundred years ago, Helland-Hansen and Nansen (1909) were able to give a detailed description of the current systems in the Nordic Seas because they observed that the currents are steered by the topography of the ocean bottom.

Recently, it has been shown that topographically steered ocean currents prefer to flow in the direction with shallow water on the right hand side (Nøst et al. 2008, LaCasce et al. 2008). Flow with shallow water on the left often becomes unstable and breaks up into small scale features with circulation crossing the depth contours. These small scale features may change the large-scale flow pattern completely. An example of this is shown in Figure 3, which shows simulations of a rotating tank experiment done by an NPI scientist at SINTEF in Trondheim. Flow with shallow water on the right (Figures 3b, c) closely follows the contours of constant depth, leading to a circulation di-

![Figure 1](image1.png)

**Figure 1.** The circulation in a rotating tank consists of vertical columns of water that move horizontally while standing upright. A column of water behaves much like a spinning top, which will also stand upright due to the rotation.

![Figure 2](image2.png)

**Figure 2** a) The surface temperature distribution in a circular ocean basin. The temperature and circulation in the basin occurs on small and large spatial scales. The small scale (L1) motion is associated with the eddies, while the large-scale motion (L2) is associated with the temperature contrast between the boundary and the center of the basin. b) A satellite image showing sea surface temperatures west of Svalbard. We clearly see eddy structures similar to the structures in a). Light colors are warm temperatures and dark colors a cold. The light blue color shows the ice distribution.
rected the same way in the two sub-basins. For the flow with shallow water on the left (Figures 3d, e) the flow in the two sub-basins are in opposite direction. This is due to the motion on smaller scale crossing the depth contours in the area marked by a black rectangle. This process is an important element in the world ocean circulation. It explains the separation of the Gulf Stream from the continental slope off USA, which has consequences for the large-scale circulation in the North Atlantic. In order to include this effect in a numerical ocean model, the continental slope region must be well resolved.

Mesoscale eddies and large-scale circulation
Mesoscale eddies have been recognized to be an important component in the poleward heat transport and overturning in the Southern ocean. At NPI, increasing effort has been put into understanding the role of mesoscale eddies in the Nordic Seas and Arctic Ocean. This is done by idealized studies and by analyzing observations from Arctic regions. The simulated eddies seen in Figure 2a transport warm water from the rim current to the center of the basin, and they are important for the sinking of cold water. The presence of similar eddies in the ocean is clearly visible in the satellite image shown in Figure 2b, showing the sea surface temperature west of Svalbard. It is likely that mesoscale eddies are important for the sinking of dense water in the Nordic Seas (Aaboe, 2008). This makes eddies an important component in the large-scale Atlantic overturning circulation.

The warmest surface temperatures shown in Figure 2b mark the Norwegian Atlantic Current (NwAC). NwAC flows northward along the continental slope off Norway and Svalbard before it enters the Arctic Ocean. This large-scale flow follows closely the continental slope, but the warm Atlantic water is also transported out into the deeper ocean as well as onto the continental shelves. This cross-slope transport is partly due to eddies. Ongoing work at NPI suggests that the eddy transport of warm water onto the continental shelf is an important component of the processes causing the abnormal ice-free conditions which have been dominating the Svalbard fjords in recent years.

Research approach
The main goal of our research is to understand the lowest order balances controlling the dynamics of the ocean currents. We have chosen not to do large-scale realistic simulations of the ocean currents, but rather to focus on the processes controlling the large-scale flow. To study and simulate these processes, we combine laboratory experiments (in rotating tanks), observational data, theory and simulations of the circulation in idealized basins. As explained above, we find that the first order balances often are set by the smaller scale processes such as mesoscale eddies (Figure 2) and cross-slope flow (Figure 3). We need to reach a higher understanding of these processes in order to improve models of the global ocean circulation and climate.

References:


Figure 3. Simulation of rotating tank experiments. a) Water depth in the tank b) and c) show velocity and surface elevation for flow with shallow water to the right (the flow will follow contours of surface elevation which can be seen as a streamlines for the flow). d) and e) show velocity and surface elevation for flow with shallow water to the left. The black frame marks the location where the flow does not follow constant depth, but instead separates from the area of sloping bottom.

More information:
- http://oceanography.npolar.no/oceanography/
- http://paoc.mit.edu/labweb/lab1/gfd_1.htm
Computer simulation of genome expression and regulation – from yeast cell division to human cancer

The study of genome expression and regulation via computer simulation may change the way molecular biologists do research in the future. Building predictive models or mathematical models to simulate the biological processes in a cell may be a first step towards a revolution in molecular biology. This will make a tremendous contribution to our understanding of human life and for example to the design of molecular targeted cancer therapies.

A U T H O R

Junbai Wang
Division of Pathology
The Norwegian Radium Hospital, Rikshospitalet
University Hospital

Genome expression and regulation
Gene expression is the process by which the genetic code of a gene is used to direct the production of the functional part of a gene: the protein. In general, the process of gene expression involves several steps, including transcription (DNA→RNA, the production of different types of RNA, including so-called messenger RNA), translation (RNA→protein, the use of messenger RNA to direct protein synthesis) and the post-translational modification of the protein. Gene regulation is a label for the cellular processes that control the rate and manner of gene expression and hence the quantity and quality of the protein.

In our work, we are interested in the transcriptional level of gene regulation, i.e., how a cell controls the function of a protein by modulating the timing, location and amount of gene expression. Gene transcription is usually controlled by interactions between regulatory factors and a regulatory region (Figure 1) located mostly on the up-stream of the transcription starting site. This regulatory region contains a short DNA sequence where the gene regulatory proteins bind to in order to activate/repress the gene expression. Here, transcription factors (TFs) are proteins that play a role in regulating the transcription of genes.

Current high-throughput technologies, for example, microarray experiments and deep sequencing methods, have allowed us to measure genome-wide gene regulation under various biochemical or physiological conditions. In the past, the design of sophisticated data mining methods and mathematical models to analyze the high-throughput measurements was very popular. Recently, the interest is shifting to build predictive models for genome

Figure 1. A simple structure of a eukaryotic gene. The regulatory region contains transcription factor (TFs) binding site (box) and nucleosome (ellipse) where some of TFs binding sites are bound by TFs (circle) but others are occupied by nucleosome. Complex interactions between TFs and other protein complexes such as histone modifying complex (triangle) and chromatin remodeling complex (hexagon) control gene regulation process for a living cell.
regulation studies by integrating heterogeneous functional genomic datasets: for example, to uncover the rules that govern the combinatorial regulations of transcription factors with the relevant genomic sequence properties.

In our present research, we are interested in developing mathematical models to identify conditional-dependent or tissue-specific transcription regulations for both yeast and humans. In particular, the computational methods that we developed may be utilized to detect disease-specific transcription regulations in human cancer, which may provide valuable information for further experimental validation and allow researchers to identify potential molecular targets for future therapy.

In silico study of gene regulation in the yeast cell-cycle

Yeast Saccharomycyes cerevisiae is a unicellular organism with a simple genomic structure and is an ideal model organism to develop and verify mathematical models for gene regulation studies. We designed a new computational framework for identifying the combinatorial control of gene expression in yeast which includes estimating protein activities using a simple linear mathematical model, summarizing high-dimensional microarray gene expression profiles into low-dimensional gene batteries (i.e., a collection of related genes expressed in unison) and the prediction of cooperative gene regulation by using various probabilistic graphical models such as Gaussian Graphical Models and Mixed Graphical Models.

The new framework was tested on datasets that describe the yeast cell-cycle (i.e., the series of events that take place in a eukaryotic cell leading to its replication), which gave promising results on inferring both TF-TF and TF-gene interactions (Figure 2) in the yeast cell-cycle. Most of our predictions are supported by literature evidence. In particular, the proposed predictive model is able to reconstruct mRNA expression of each gene battery by only considering DNA sequence properties of genes and activity profiles of TFs. The result demonstrates that biological processes may also obey certain law or code, such as Ohm’s law in physics, to control genome regulations. Besides predicting the gene expression level and protein activity profiles, the in silico study of gene expression also discovered several biologically interesting observations in the yeast cell-cycle. For example, negative correlation of protein-protein interactions, low affinity of protein-DNA interactions, and noises in gene expression may play pivotal roles in the yeast cell-cycle as well. The new framework may easily be modified to the study regulation of gene expression in higher eukaryotes.

In silico study of human p53 target genes

Human p53 is a “master” protein which responds to a variety of intrinsic and extrinsic stress signals, activates the expression of its downstream genes and controls a wide range of cellular processes. The p53 regulation may be a perfect system to validate the new computational inference method in investigating human gene regulation. To study the regulatory functions of p53, we
need to understand transcriptional target genes of p53 since these are the cornerstones to elucidate the complex regulatory mechanisms of the p53 protein. In particular, the identification of commonly supposed (putative) p53 target genes will contribute tremendously to the development of therapeutic strategies for treating cancers.

Therefore, our present in silico study of human gene regulation is aimed at inferring dynamic regulations by the human p53 protein by using various functional genomic datasets with a novel nonlinear mathematical model to describe transcriptional regulatory systems. The proposed mathematical model considered the positive (negative) regulation, the number of protein binding sites, the cooperative binding of TFs, the protein binding structure of the p53 transcription factor and the effect of time delay on gene regulation. The kinetic rates and all model parameters of present quantitative modelling of p53 regulation are estimated by a genetic algorithm. The method can be used to estimate not only the regulation relationship between TF and its downstream genes but also the upstream activities based on the expression levels of target genes.

An initial test of the new quantitative model on available p53 microarray data provides promising results for inferring dynamic regulation of the tumour suppressor gene p53. For example, the mathematical model is able to reconstruct the dynamic expression profiles of p53 target genes (Figure 3). Our predicted top 100 putative p53 target genes have at least two perfect matches of p53 consensus binding motifs on the upstream regions (~80% of targets); and we suggest cooperative regulations (i.e., either direct protein-protein interaction or indirect protein-DNA interaction) may be essential for p53 to regulate functionally diverse target genes. The development of a quantitative model for inferring dynamic gene regulation will not only generate useful hypothesis for the biologist to perform further wet lab analysis (involving experiments handling biological material that require specialized utilities), but also significantly reduces the potential cost for performing large scale genome-wide studies.

From these studies, we have learned that a new door - in silico studying of genome regulation - is opened to the wet lab scientists which may change the way people do molecular biology research in future.

Building predictive models or mathematical models to simulate the biological processes, may be a first step towards a revolution of the molecular biology field. This will make a tremendous contribution to our understanding of human life and to the investigation of potential biomarkers for disease in order to design specific molecular targeted therapies.

Acknowledgements
This work relies on the use of the cluster facilities of the University of Oslo and the Notur project. The present work receives support from the Norwegian Cancer Society and some financial support from the Ovarian Cancer Foundation of the Norwegian Radium Hospital. The work is the result of inspiring collaborations with Drs. Ben Davidson, Ola Myklebost, Harmen Bussemaker, Jahn M. Nesland, Jan Delabie, and in particular Dr. Tian Tianhai in the p53 project. The author also appreciates the improvements to this article that were suggested by Dr. Morigen. Finally, the work described in this article is published in a number of scientific publications that can be obtained from the author.

Figure 3. Comparison between reconstructed expression profiles by our quantitative model (line) and measured expression activities by microarray experiments (stars). A and B: two genes are positively regulated by p53. C and D: two genes are negatively regulated by p53. All four genes have small model estimation errors. E and F: two genes have relatively large model estimation errors but the simulated expression profiles still match the experimental data quite well. The relatively large errors may be caused by low gene expression or noise in gene expression.
Alloy design

SINTEF’s price for excellent research in 2007 was rewarded to Sigmund Andersen and Calin Marioara from SINTEF Materials and Chemistry for their work on characterization and properties of nanoparticles precipitated inside aluminum alloys during aging. In this project, experimental techniques are used in combination with simulations for which high performance computing has regularly been used.

Norway produces over one million tons of aluminum alloys (most of it is Al-Mg-Si) per year, and the industry is eager to understand more of the fundamental physics in this field. This is the background for these Al-projects, which over many years have been collaborations between SINTEF, NTNU and Norwegian light metal industry, with Hydro as the main industrial partner. Physical properties, such as strength and ductility, are given as a function of the materials structure at the micro- and nanoscale. When a typical aluminum alloy, consisting of ~98% Al and ~2% alloying elements (e.g., silicon, magnesium), is heat treated at for example 180°C for 5 hours, the strength of the material can increase by a factor 3 or more! This is caused by the formation of small (nanometer scale) hardening precipitates inside the alloy bulk material. These precipitates consist of Si and Mg (and Al) and help to prevent further deformation. Figure 1 shows the different length scales involved. The form, structure and strengthening properties of these particles depend on the alloy composition and the thermo-mechanical history of the material. Being able to determine all these parameters and understand how they affect the material’s physical properties, the right alloy for the right applications can be designed – and ‘alloy design’ can be used to tailor materials with the wanted properties.

Precipitation, growth, and transformation in these alloys are inherently complex. Especially the finest precipitate structures formed at the early stages are difficult to characterize. Several mechanisms are at work simultaneously. This makes the calculation from first principles necessary to understand the processes.

The general procedure to determine the crystal structure of the small precipitates [see references 2 and 3 and references therein] in this project is illustrated in Figure 2. First, several advanced experimental techniques are used in a more qualitative way. High Resolution Transmission Electron Microscopy is used to estimate positions of columns within the crystal structure’s unit cell. The composition of the precipitates can be measured using either Energy Dispersive X-ray analysis or Atom Probe Tomography. Electron Diffraction data, recorded from different orientations of the precipitates, is analyzed qualitatively.

The experimental data, combined with knowledge of similar structures, is used to suggest a trial model for the crystal struc-
ture. Quantitative analysis of the diffraction data is then applied to refine the trial model. Here, the experimentally measured diffracted intensities are compared to simulated intensities based on the model and the model parameters are refined to minimize the difference between experimental and simulated diffraction.

Density functional theory (DFT) calculations are an essential tool here. In this project, the VASP \cite{Kresse1996} DFT software package is used to verify the experimentally obtained structure, by relaxing the structure energetically. For a reliable structure, obviously the relaxed structure should not differ significantly from the original one. Also, in some cases it may not be possible to distinguish between 2 (or in some cases more) ‘candidate’ structures based solely upon the experimental data. Energy calculations on these different structures can then be used to determine the most likely structure. Recently, VASP calculations have been started which include the precipitate-matrix interface. Especially for very small structures, the interface energy contribution can be very significant, and needs to be accounted for.

The calculation time in VASP scales approximately with the third power of the system size, and the required computational resources (both in terms of CPU hours and memory size) will easily exceed the specifications of a single PC, making the availability of a supercomputing facility essential to this project.

References:
Processing and distributing language data
The amount of digitally available language data is phenomenal and growing. Entire libraries are being scanned, the number of blogging and news sites is exploding, Wikipedia by far surpasses printed encyclopedias in size, and companies around the world are going paperless. The cornucopia of information at our fingertips means, however, that we are forced to search for the proverbial needle in the haystack.

The current generation of search engines looks only for keywords, does not distinguish between different meanings of a word, fails to recognize semantic relations and cannot be queried from a different language. Searching for *all athletes who lost their lives due to a disease last year* will not be optimal if the keywords are matched separately and literally, and the semantic relations between the concepts in this query are not taken into account. There is no good chance of matching this query with *In 2007, the marathon runner died of non-Hodgkin’s lymphoma*. It is also wishful thinking to phrase the search in Norwegian and expect the answer to include information found in several other languages.

Advanced information retrieval tasks require a deep analysis of structure and meaning in written and spoken language, in addition to indexing and classification. Providing information in useful ways may also require other linguistic processing such as summarization, paraphrasing, translation, text-to-speech and speech-to-text conversion. For about half a century, researchers in computational linguistics and language technology have been slowly chipping away at the challenges.

From an information theoretic viewpoint, language is a code to be cracked. Some ancient scripts have in fact never been fully deciphered. But even common present-day language can only be understood in a context of sufficiently rich background knowledge. The meaning of the common phrase *Christmas tree* may be very different on an oil field, in drag racing, on the British railways, and in test taking (the curious reader will find the definitions in Wikipedia). The famous Groucho Marx quote *One morning, I shot an elephant in my pajamas. How he got in my pajamas, I don’t know* illustrates unexpected structural ambiguity, which is in fact rampant in natural language. Combinatory explosions of ambiguities may easily yield hundreds or thousands of analyses for a single sentence of some length. Humans can usually identify the intended reading in a fraction of a second, although there are exceptions that can lead to either irritating or hilarious misunderstandings.

Symbolic models of natural language processing typically involve parsers and automata as well as logic and inferencing. The acquisition of symbolic knowledge, however, has proved to be a bottleneck for these models. A fair amount of world knowledge is necessary to correctly interpret the notice *Your second hand and body towel are in the wardrobe found in an Istanbul hotel room*. An alternative approach to symbolic models relies instead on statistical approaches to harvest distinctions, preferences and correspondences from large quantities of speech or text, in the order of millions or even billions of words. The empirical data needed for statistical language modeling are generically called *language resources* and include marked up corpora of text and speech, as well as derived data such as lexicons, thesauri, word nets, frequency lists, and automatically induced grammars. While some classification tasks may take raw language data as input, of which there is plenty on the Web, most natural language processing tasks are crucially dependent on corpora that are previously analyzed, disambiguated and annotated with detailed linguistic information.

A typical method for obtaining the most probable analysis of a sentence uses stochastic principles to prefer certain rules or readings over others in a given context. Stochastic parsers need to be trained on previously analyzed and disambiguated corpora, commonly called treebanks, which are essentially large databases of sentence analyses at the syntactic and sometimes at the semantic level. The secondary data generated during treebank construction may be hundreds or thousands of times bigger than the primary data.

For machine translation, a typical statistical approach relies on the frequency based matching of *n*-grams [sequences of *n* words] in the source language with known translations in the target language. This approach requires access to a large parallel corpus, aligned at word level. Current models have gone up to using 5-grams, more than a billion of which occur with some reasonable frequency in English on the Web. Parallel corpora are also used to study translational correspondences between words in dif-
different languages, which in turn provide a basis for building ontologies, thesauri and translation dictionaries.

The accessibility of suitably annotated and coded language resources is a prerequisite for developing statistical language models. Since it is difficult and expensive to hand-code large enough corpora, there is essentially a bootstrapping problem. The quality of large scale natural language processing systems depends on the quality and size of the available data that previously have been processed and annotated by humans or computers (or, more likely, in a computer-assisted manual procedure). Bottlenecks occur when there is a lack of suitable computer tools or trained human workers for the corpus building tasks, which need to be repeated for every new language (although if we wait long enough, many languages will be extinct before we start working on them). Other bottlenecks may consist of copyright problems, different coding standards, or other incompatibilities. For those reasons, it is not sufficient to produce good language resources, but it is equally important to build the infrastructures that get them out to the research and development community.

A current large-scale European effort to promote the accessibility of language resources is the CLARIN project. This project aims at providing a comprehensive data service to the humanities with an emphasis on written and spoken language. Resource and service centers will be connected via grid technology to form an integrated domain. National IT centers such as CSC in Finland are already highly involved in this project. Differences in format, structure and terminology in the data sets must be overcome, but ultimately, the resources should be stable and widely available as well as persistent, easily accessible, user tailored and extensible.

Both the construction of language resources and their eventual use in applications involves intensive data processing. Natural languages are very complex; the currently largest computational grammar for Norwegian has 293,394 rule disjuncts and the average ambiguity level runs in the hundreds of analyses per sentence. The highly optimized XLE parser takes from a few hundredths of a second to over a minute, depending on the sentence, for a full multi-level parse on a modern desktop machine. Since the development of large scale grammars implies testing by regular reparsing of a treebank with new versions of the grammar, it is clear that the availability of computational power may become an issue at some point. While present-day desktops are capable of fully parsing a thousand sentence corpus in a few hours using efficient and special-purpose software, the demand for computing power obviously increases as more sophisticated models involve larger corpora, more complex grammars and more fine-grained analyses.

Several companies working on the next generation of search engines are heavily relying on linguistic tools and semantic web technologies. A San Francisco based company has already parsed all of Wikipedia and extracted all semantic relations, so as to allow the kind of non-literal matches mentioned earlier in this article. Their ultimate goal is to parse everything on the Web, which will generate petabytes of secondary data. The management of computer power to be allocated to the various language analysis tasks turned out to be one of the biggest challenges in their project. They are now running large computer centers (somewhat similar to Google’s) with data flows optimized for linguistic tasks and information retrieval.

Also in other humanities fields such as literary computing, history, etc., some advanced models may require significant computational efforts. Stylometry and author attribution, for instance, have progressed from using simple frequency distributions to more complicated models involving memory based shallow parsing, genetic optimization of feature selection and classifier building, with various optimizations and compressions. Especially genetic optimization has contributed to a boost in accuracy up to about 80%, but these kinds of algorithms are known to be computationally expensive. Other computationally demanding methods in text studies include Monte Carlo methods and neural networks; the latter have for instance been applied to the works of Marlowe and Shakespeare.

Language processing has a long tradition of pursuing ambitious goals. Already in the 1950s, when Father Busa turned to IBM for help in producing a concordance of the complete works of Thomas Aquinas, founder Thomas J. Watson initially said it couldn’t be done on their state-of-the-art machines. Until recently, however, despite ever more ambitious goals and more demanding methods, academic research groups in language processing and other humanities computing areas were not known for knocking on the doors of high performance computing centers.

A possible reason is that linguists and other scholars who have been raised in a humanities tradition have never been encouraged to ‘think big’ with respect to research infrastructures, in contrast to the natural sciences. They have perhaps tended to dimension their research to the platforms they have immediately available and focused on theory, formalisms and small scale models, while leaving ‘real world’ deployment to large corporations who do possess ample computing resources, and importantly, the technical support staff to manage them and to provide a bridge between the needs and the solutions.

This situation seems to be changing as the sophistication of new models and the data to be processed to develop them increases sharply. The recent LOGON Norwegian-to-English machine translation project took a hybrid approach combining symbolic and statistical models, which involved very CPU-intensive tasks. Typically exploring thousands of candidate translations per input sentence, end-to-end translation of the 5000-sentence development corpus took 155 CPU hours, but a year later, when more vocabulary was added, this number went up to 246 CPU hours. Now the research team has joined forces with the University of Oslo’s high-performance computing group. These developments are interesting and should be supported by placing advanced computer techniques for large scale data processing on the PhD studies curriculum.

THE CLARIN PROJECT:

http://www.clarin.eu
Environmental impacts of hydropower regulations

The Research Council of Norway has funded a study looking at how hydropower regulations affect the geometry of the delta of the reservoir. An important example in Norway is Lake Øyern, having the largest freshwater delta in Northern Europe. The delta covers an area of about 9 km², and it is an important bird habitat and nature protection area (Fig. 1). The geometry of the reservoir is constantly changing due to erosion and sedimentation processes. Many plants, insects, birds, and animals living in the delta are affected by the size of wetlands with favorable water depths. The reservoir is regulated with hydropower plant Solbergfoss, affecting the water levels. From an environmental point of view, it is therefore important to assess how the regulations affect the changes in the delta geometry.

Figure 2 shows the resulting water velocities from a computation with a numerical grid covering the whole delta. The grid is three-dimensional, and has varying number of grid cells depending on the water depth. In areas where the river bed is above the water surface, no cells are generated. The gridding algorithms capture the complexity of the delta geometry. For this particular grid, maximum number of cells in the vertical direction is 15, and the total number of cells is 500,000. Figure 3 shows how the grid resolution affects how the details of the geometry are resolved.

The steady computation producing the results in Fig. 2 was run on the Njord cluster at NTNU, taking about 2 hours on a 16 CPU
The program was parallelized with OpenMP. Later studies will include time-dependent evolution of the delta, and then the computational time will increase with several orders of magnitude. The use of Njord will then be necessary for obtaining a result.

Movement of sand dunes

The computation shown in Fig. 2 includes many simplifications and different subprocesses. One such process is the movement of sand dunes on the river bed. Looking at Fig. 2 in detail, there are some irregular variations in the velocities for some locations. The reason is the existence of large sand dunes on the bed, causing the water depth and the resulting water velocity to change. The sand dunes affect the roughness of the bed and thereby the water velocity and water surface slope. It also affects the bed shear stress and the sediment transport capacity. Therefore, a separate study of the dunes is included in our studies. In the study, one section of a delta branch was modelled in more detail. Fig. 4 shows a 3D view from this computation, where the bed surface is given with the colours showing the vertical elevation. A longitudinal section of the geometry is also shown, where the colours give the computed water velocities. The movements of the dunes were computed and compared with side-scan sonar measurements at two different time intervals. Fairly good agreement was found. The computation was done on the Njord cluster at NTNU.

Turbulence and detailed flow fields in bends

The main changes in the largest channels of the delta are caused by erosion of the side banks. This lateral erosion takes place at the outside of river bends due to secondary currents and turbulence in the river. The secondary currents sweep the sand from the outside to the inside of the curve. To predict the process, it is necessary to be able to compute the secondary currents and the associated turbulence field. The accuracy of the sediment computations will therefore be affected by the accuracy of the turbulence model. Another project funded by the Research Council of Norway looks at improvements of turbulence models for river computations. The turbulence models are tested on simpler channel geometries, as shown in Fig. 5a. This is a small laboratory flume with a meandering channel. The steady turbulence model used in Fig. 2 and 4 is compared with a more advanced method, called large eddy simulation (LES), where the largest turbulent eddies are directly resolved in space and time. The instantaneous velocity field showing the turbulent eddies in a cross-section of the bend is given in Fig. 5b. The time-averaged values in the same cross-section are shown in Fig. 5c. LES is considered to be at the forefront in the science of turbulence modeling, and it requires enormous computational resources. The grid used in the LES of the meander employs more than 12 million cells. The computation used more than 40 000 CPU hours on the Stallco cluster in Tromsø. Without a high-performance computing cluster, this calculation would not have been possible.

Further work

Our group has done some initial work on modeling how vegetation affects the flow field and sediment deposition. The delta of Lake Øyern received sediment deposits during the flood in 1995, and this was recorded in detail by the Norwegian River and Electricity Authority (NVE). The plan is to model the sediment deposition during the 1995 flood some time this fall. The water flow field was affected by the submerged vegetation during the flood. The plan is to do further time-dependent analysis of the geomorphological changes in the delta.
Another ongoing project funded by the Research Council of Norway is to look at the computation of damages from local scour, often occurring during floods. Fig. 6 shows the computed erosion around a cylinder placed in a sandfilled flume. Originally, the bed was flat, but the computed currents and turbulence around the cylinder caused scour to form. The case is a simplified situation of flow around a bridge pier. When the scour hole exceeds a given depth, the pier will collapse. 60% of all bridge collapses in the USA are caused by this mechanism. The case shown in Fig. 6 was computed on a workstation with four processors. However, more complex natural geometries will require much finer grids, and then it will be necessary to use high-performance clusters also for these computations.

Figure 5. a) 3D view of the computational domain of the meandering channel
b) Cross-section at the apex, where the colours show the instantaneous water velocity
c) Cross-section at the apex, where the colours show the time-averaged water velocity.

Figure 6. 3D view of the local scour hole around the pier. The colours show the bed elevation. Water is flowing from the left to the right.

Conclusions
The high-performance clusters in Norway enable our research group to stay in the international scientific forefront on computational modeling of rivers and sediment transport. The research carried out using the clusters progress the science of numerical modelling of erosion damages from floods. It also benefits predictions of environmental effects from regulation of hydropower plants. Our group will need increased computational resources in the future.
A Resource Allocation Committee is appointed by the Norwegian Research Council to allocate computing time on Notur’s High Performance Computing (HPC) resources. The allocation is based on applications from the research groups and decisions are made based on the quality of the research and the need for compute resources. This article gives a brief overview over the resources available, the allocation process and how the Committee tries to balance multiple and sometimes conflicting goals. Hopefully, this can assist you in writing better applications.

**Compute time**
The allocatable resource on the HPC facilities is first and foremost compute time (CPU cycles). While the overall goal of allocating resources based on quality of research and needs sounds obvious and fairly easy to deal with, things get muddled up in practice. Additional criteria and considerations must be taken into account.

**Simplicity**
Time is one of the scarcest resources for any scientist, thus we want the application procedure to be simple and not time consuming. However, the application must contain sufficient information to allow the Resource Allocation Committee to make qualified decisions. Obviously, wrong decisions made for large requests for resources pose bigger problems than mistakes made for smaller ones. Thus, more detailed information and a more thorough evaluation is needed for large applications for resources.

We divide applications into two categories. For large requests for resources, we require a thorough description of the project, the computing needs and demonstrated performance of the software. These applications go through a peer-review process involving international experts.

It is not cost-effective to invest a similar effort for small requests for resources, neither for the applicant nor for the Committee. Thus for smaller applications, a shorter project description is sufficient plus some characteristics of the computing profile.

The borderline between “large” and “small” requests for resources is a moving target. It needs to be adjusted (increased) regularly as many-core systems are being introduced, the total compute capacity grows and the cost of CPU-cycles decreases.

**Global versus local optimisation**
When writing an application for resources, a researcher has only knowledge of his/her own personal compute needs and maps these as good as possible onto the systems available in the Notur project. The researcher can’t possibly know what other researchers apply for and therefore does not have a “global picture”. When all requests for resources are on the table, the typical situation is that these are far from evenly distributed across the different systems.

An important task for the Resource Allocation Committee is therefore to perform a “global optimisation” step to get good utilization of all the systems, yet ensuring that the user requests and software applications are mapped onto suitable systems. However, situations may occur where users are moved to systems they consider suboptimal from their local viewpoint.

**Flexibility**
Genuine research implies entering into the unknown, thus it is hardly ever possible to give a precise estimate on the amount of resources one needs, and when these are needed. Many researchers underestimate the time it takes to develop and verify new software and subsequently experience delays. For that reason, we frequently observe projects that only consume a small fraction of a large allocation. At the next round for applications, they apply for a continuation of the project, adding a sentence like “Sorry, we are delayed”. To compensate for such inefficient use of allocations, the committee overbooks the systems, i.e., more CPU-hours are allocated to projects than there are actually available on the system. History shows that overbooking by 40% works remarkably well.

The opposite may also occur. Initial computations may reveal that further or larger studies are needed and more computational resources are required. If that is not possible within the current allocation, it is possible to apply for extra CPU-hours. Such requests are accommodated depending on availability.

**Application strategies**
One popular strategy that does not work is asking for X times the amount one actually needs, assuming the total available allocations will be oversubscribed by a factor X, and assuming that the Committee handles oversubscription by a flat cut across all applications (by a factor X). Oversubscription of available resources is quite common. In such cases, the Committee looks at the track record of the projects in their
The resource Allocation Process

use of previous allocations. Projects that repeatedly ask for much more than they use without proper explanation, are eventually reduced in size accordingly.

The next thing the Committee does is looking for projects with flexibility, i.e., projects that allow that they are moved to compute platforms where there is no or little over-subscription. Being moved to a platform with available capacity increases the possibility for a project to get granted what is asked for. Flagging willingness and options to use other platforms is a good strategy in writing applications.

Only when these two possibilities have been explored, the painful cutting starts...

Advanced user support
Since a couple of years, there has been an advanced user support program that allows researchers to apply for higher level application support, e.g., to tune performance, parallelize code or enable complex applications. This has been quite successful. I expect the importance of this program to increase, and I would urge experienced as well as beginning users of the HPC facilities to take a careful look at this and see if it is of any interest and how to exploit it. A basic requirement for advanced user support to work is that an inventory is made of what is needed for a code to become an efficient scalable parallel code, and a step-by-step plan for how to achieve this.

The reason why this will be more important in the future is that we are moving away from a situation where having access to sufficient compute cycles is the sole obstacle to satisfy computational needs, to a situation where limitations in software become the real bottleneck.

Last year has been exceptional in the history of HPC in Norway. Thanks to extra funding from the Research Council of Norway, it has been possible to install new large facilities that extended the computational capacity by more than a factor of 20 in terms of CPU-cycles. This provides a unique opportunity for Norwegian computational scientists to undertake simulations that were previously beyond reach. However, it is not simply a question of “pushing the pedal to the floor” and let this new wonderful hardware produce the long awaited spectacular results.

The twist is that the increase in compute power comes in the form of increased numbers of CPUs (or more precisely cores), not by increasing the compute power of the individual CPUs, as has been the case for new facilities in the past decade. The implication is that only those researchers that have scalable parallel codes that can use many processor cores simultaneously, are positioned to immediately profit from the increased computational capacity. Moreover, the new systems are distributed memory systems, consisting of many compute nodes, each with few cores that share memory. Currently, at most 16 cores have access to shared memory. To be able to use hundreds or thousands of cores simultaneously, parallel code must not only be scalable, it also needs to be parallelized using a distributed memory programming paradigm (in practice: MPI-code).

In my crystal ball, I see this trend continue in the next 3-5 years (= the next renewal cycle). This implies that the next generation systems also will achieve their increase in speed mainly by adding more cores, not by squeezing the clock frequency. We may see some increase in the number of cores sharing a common address space, but large numbers of cores available to a single program will only be possible through distributed memory. Thus the need for scalable MPI-code will be even more important in the near future, and important CPU-hungry codes not satisfying this requirement are facing a bleak future! It may now be just as important to invest in rewriting application code as it is to invest in faster hardware. Notur and the Resource Allocation Committee are prepared to assist in this effort through the program for advanced user support.

Storage archive
A new allocatable resource is currently being added to the national e-infrastructure in Norway, namely storage. The new storage facilities in question are primarily used for archival (long-term) storage but also to host actively used scientific data collections. From an allocation perspective, storage is completely different from CPU. CPU-cycles are extremely short-lived and have to be used there and then. The very idea of storage is the opposite. It does not make sense to allocate long-term storage to a research project as a succession of short term allocations; it must be dealt with using longer time horizons. But allocations are not necessarily indefinite. This requires well-defined mechanisms, responsibilities and criteria to decide when data becomes obsolete. Here, we need to step carefully, deleted data may not easily be recovered.

On-going research projects must be allowed to grow their data sets, and projects must have access to permanent as well as temporary storage when needed. Over-booking storage capacity is not necessarily a good strategy for optimal utilisation in the case of archive storage, and high usage (filled disks) is probably not a good criterion for efficient use.

The new NorStore project will be given the responsibility for providing this service and the Resource Allocation Committee will be responsible for the allocation process.

Grid
For transparent and easy access to distributed HPC and storage systems, grid application software should be used. Over the last decade, there has been a frenetic activity all over the world to create grid standards and middleware. Now is probably the time to make the hard choices and implement the grid-application software that will ease the use and increase the productivity in the Norwegian system. (Sorry folks, no unique universal standard or middleware this time either).

This has to be done in collaboration between the different user groups who know their scientific needs, and the technical expertise available at the HPC centres, to determine the possibilities and limitations for building application grids. Resources for such work will be allocated based on applications, and again, the Resource Allocation Committee will decide.
NOTUR2008

NOTUR2008 was the seventh edition of the annual meeting on High Performance Computing and Infrastructure for computational science in Norway. The NOTUR2008 conference focused on how one can translate the increase in compute capacity in Norway into increased scientific production and how a wide range of sciences can make use of this capacity. The conference also provided updates on grid and storage initiatives. The conference was held June 3-5 at the University of Tromsø.

Notur User Survey 2008

The results from the survey that was carried out in April 2008 among the user community of the High Performance Computing facilities of the Notur project are now available on www.notur.no/publications/

PRACE survey

The Partnership for Advanced Computing in Europe (PRACE) prepares the creation of a persistent pan-European HPC service, consisting of several centres providing European researchers with access to Petascale computers that form the top level of the European HPC ecosystem. PRACE is a project funded in part by the EU’s 7th Framework Programme. Recently, PRACE conducted a number of surveys among the users with the largest computational needs within the countries that participate in the consortium. These include a comprehensive survey of HPC training and education needs and a survey of the utilisation of HPC resources across Europe. The results of these surveys and all the latest news related to the PRACE project can be found on website: www.prace-project.eu/documents/

eVITA winter school, 11-16 January 2009, Geilo

The eVITA winter school is an annual meeting series where researchers meet to collaborate, exchange ideas and get updated on developments within eScience. The winter school primarily aims at graduate, doctoral, and postdoctoral students, but it is also open for advanced undergraduate students and senior researchers. The focus of the winter school is on new methods and theories within mathematics, statistics and computer science, but with an eye to applications of national interest.

Further announcements and programme will become available on www.sintef.no/evita

Subscription: If you would like to subscribe to the paper version of the META magazine, please go to www.notur.no/publications/magazine/

The Notur II project provides the infrastructure for computational science in Norway. The infrastructure serves individuals and groups involved in education and research at Norwegian universities, colleges and research institutes, operational forecasting and research at the Meteorological Institute, and other groups who contribute to the funding of the project. Consortium partners are UNINETT Sigma AS, the Norwegian University of Science and Technology (NTNU), the University of Bergen (UiB), the University of Oslo (UiO), the University of Tromsø (UiT), and the Meteorological Institute (met.no). The project has a 10-year duration (2005-2014). The project is funded in part by the Research Council of Norway (through the eVITA programme) and in part by the consortium partners.