Computational e-Science: Studying complex systems in silico

A National Coordinated Initiative : 'White Paper'

Sloot, P.M.A.; Frenkel, D.; van der Vorst, H.A.; van Kampen, A.; Bal, H.; Klint, P.; Mattheij, R.M.M.; van Wijk, J.; Schaye, J.; Langevelde, H.-J.; Bisseling, R.H.; Smit, B.; Valenteyn, E.; Sips, H.; Roerdink, J.B.T.M.; Langedoen, K.G.

Publication date
2007

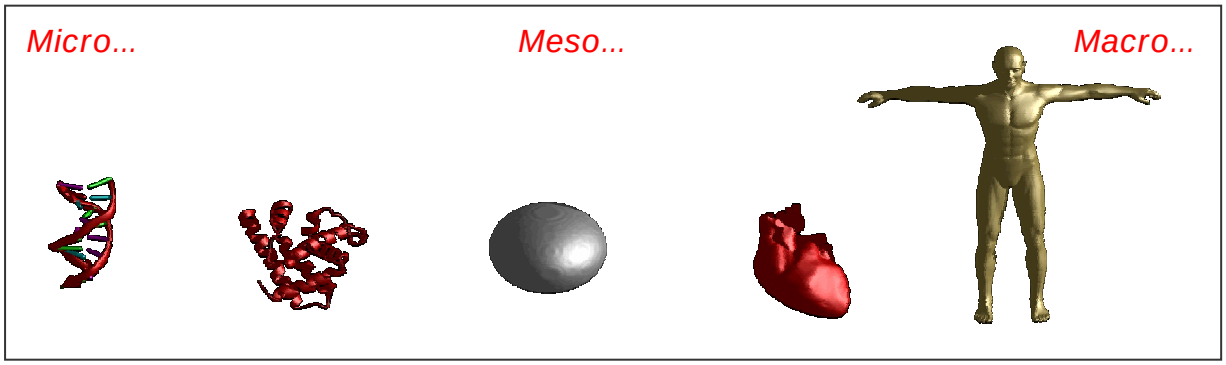
Document Version
Final published version

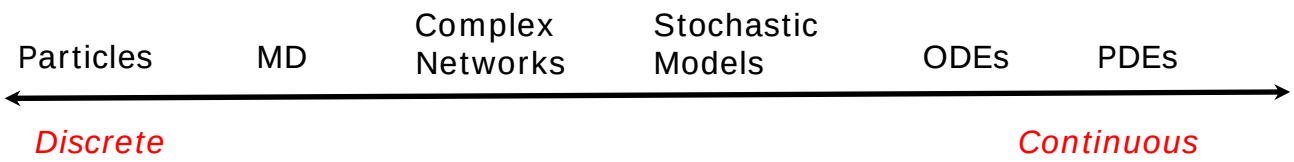
Computational e-Science

Studying complex systems *in silico*

A National Coordinated Initiative

'White Paper'





February 2007

P.M.A. Sloot, D. Frenkel, H.A. Van der Vorst, A. van Kampen, H. Bal, P. Klint, R.M.M. Mattheij, J. van Wijk, J. Schaye, H-J Langevelde, R. H. Bisseling, , B. Smit, E. Valenteyn, H. Sips, J.B.T.M. Roerdink, K.G. Langedoen.

Outline

Executive Summary

The vision of the *National Coordinated Initiative* 'Computational e-Science' is to advance innovative, interdisciplinary research where complex multi-scale, multi-domain problems in science and engineering are solved on distributed systems, integrating sophisticated numerical methods, computation, data, networks, and novel devices.

We aim to become one of the world-leaders in the field of computational e-science. We will build on the impressive results obtained from previous NWO programs and the (new) Dutch infrastructural programs that focused on understanding through modeling and simulation of interdisciplinary, multi-domain, multi-scale processes in science and engineering.

Computational e-science enhances theoretical and experimental progress in many areas of science critical to the scientific and societal needs of the 21st century. It is the field of study concerned with constructing mathematical models and numerical techniques and using computers to analyze and solve complex scientific and engineering problems. Successes have been documented in areas such as biotechnology (e.g., genomics, cellular dynamics), advanced energy systems (e.g., fuel cells, fusion, alternative energy), nanotechnology (e.g., sensors, storage devices), and environmental modeling (e.g., climate prediction, pollution remediation, flood prediction). Computational science offers the best near-term hope for progress in answering many scientific questions in such disparate areas as the fundamental structure of matter, the functions of enzymes, the production of heavy elements in supernovae and the spread of infectious diseases.

This National Coordinated Initiative will bring together Dutch top experts in the fields of computer science, numerical mathematics, astrophysics and astronomy, bioinformatics, physics and chemistry.

The perspective of *Computational e-Science*

Understanding complex systems

Recent advances in experimental techniques such as detectors, sensors, and scanners have opened up new windows into physical and biological processes on many levels of detail. The resulting data explosion requires sophisticated techniques, like grid computing and collaborative virtual laboratories, to register, transport, store, manipulate, and share the data. The complete cascade from the individual components to the fully integrated multi-science systems crosses many orders of magnitude in temporal and spatial scales. The challenge is to study not only the fundamental processes on all these separate scales, but also their mutual coupling through the scales in the overall system, and the resulting emergent properties. These complex systems display endless signatures of order, disorder, self-organization and self-annihilation. Understanding, quantifying and handling this complexity is one of the biggest scientific challenges of our time [*Barabasi 2005*,3].
A prototypical example comes from biomedicine, where we have data from virtually all levels between 'molecule and man' and yet we have no models where we can study these processes as a whole. It is a real complex system: from a biological cell, made of thousands of different molecules that work together, to billions of cells that build our tissue, organs and immune system, to our society, six billion unique interacting individuals. The complete cascade from the genome, proteome, metabolome, physiome to health constitutes multi-scale, multi-science systems, and crosses many orders of magnitude in temporal and spatial scales [*Finkelstein 2004*,6]. Another example comes from flood prediction: sensor data from dikes and up-to-date river flow patterns combined with information on the state of the locks and bridges need to be integrated with satellite data and weather models to predict flooding and give decision support to local governments. For this we need to build time critical numerical models that can simulate ‘what-if’ scenario’s, where hydraulic, flow and weather prediction models are integrated.
The sheer complexity and range of spatial and temporal scales defies any existing numerical model and computational capacity. The only way out is by combining data on all levels of detail with for instance large scale particle-based, stochastic and continuous models; an open research area. The challenges include understanding how one can reconstruct multi-level systems and their dynamics through computational simulation within virtual laboratories that connect models to massive sets of heterogeneous and often incomplete data. Conceptual, theoretical and methodological foundations are necessary in understanding these multi-scale processes, dynamic networks, and the associated predictability limits of such large scale computer simulations. The Netherlands has a unique opportunity to become a world leader in this research field of Computational e-Science through the existing expertise, the international recognition and the advanced National infrastructure.

The e-science challenge

One of the big scientific challenges of the Computational eScience initiative is to demonstrate that the implementation of distributed (grid-based) virtual organizations truly facilitates and advances collaborations and scientific progress between fields that would otherwise not easily collaborate. Such virtual organizations will force scientists to start thinking about integration of research efforts with respect to knowledge, data and tools. This will likely accelerate science in ways that can now only be imagined.

The Dutch Dimension

Within the Netherlands Organization for Scientific Research (NWO), computational science has been an area of focus for many years [NWO, 20]. It started with the

programs on Massive Parallel Computing (MPR), followed by the programs Computational Science (CS) and Computational Life Sciences (CLS). Recently a series of e-science related initiatives have been implemented [GLANCE, VIEW, STARE, 12]. In addition the foundation 'Nationale Computer Facilities' (NCF) as part of NWO, is very active in the field of supercomputing, clusters and grids.

We aim to become one of the world-leaders in the field of computational e-science. We will build on the impressive results obtained from previous NWO programs and the (new) Dutch infrastructural programs, that focused on understanding through modeling and simulation of interdisciplinary, multi-domain, multi-scale processes in science and engineering.

Virtual collaborative laboratories

The radical increase in the amount of IT-generated data from physical, living and social systems brings about new challenges related to the sheer size of data, transforming science and engineering as we know it. It was this data 'deluge' that triggered the research into grid computing [Foster et al, 2002, 7], [Hey et al., 9]. Grid computing is an emerging computing model that provides the ability to share data and instruments and to perform high throughput computing by taking advantage of many networked computers able to distribute process execution across a parallel infrastructure. With the emergence of grid computing, the necessity for advanced methods to share and manipulate data, instruments, sensors and computational power became essential. This resulted in a worldwide effort to build collaborative virtual laboratories (e.g.: [Afsarmanesh et al., 2002,1], Virtual Laboratory for e-Science [Vle, 23] and LMS [LMS 2005,15]). They represent heterogeneous and distributed problem solving environments that enable groups of researchers located in different geographically spread centers to work together, sharing resources, data and computational tools. This type of research is denoted by *e-Science*. The term describes computationally and data intensive science that is carried out in highly distributed network environments requiring grid computing. The term was originally coined by John Taylor, the director general of the United Kingdom's office of science and technology and was used to describe a large world wide initiative starting in the beginning of 2000. Examples of this kind of science include social simulations, particle physics, earth sciences and bio-informatics. Figure 1 depicts a typical architecture to conduct e-Science experiments [from Sloot et al., 2006, 19].

Virtual collaborative laboratories, built on grid-based distributed resources, will provide new windows into processes on all scales and across all disciplines. We need novel computational methods to exploit the resulting information. The challenges include understanding how one can reconstruct multi-level systems and their dynamics by integrating simulation with virtual laboratories that connect models to massive sets of heterogeneous and often incomplete data.

The vision

The vision of the *National Coordinated Initiative* 'Computational e-Science' is to advance innovative, interdisciplinary research where complex multi-scale, multi-domain problems in science and engineering are solved on distributed systems, integrating sophisticated numerical methods, computation, data, networks, and novel devices.

Over the last decade great progress has been made in providing theoretical foundations and conceptual frameworks for the science of complex systems, including emergent phenomena, co-operative behavior, complex networks, adaptability, evolution, and so on. Theoretical understanding has evolved from simple toy models into structured models in which the heterogeneities and details of the system under study are more and more included. The challenges for the best prediction and predictability include

understanding how one can reconstruct multi-level systems and their dynamics by integrating simulation with data platforms that connect models to massive sets of heterogeneous and often unreliable data.

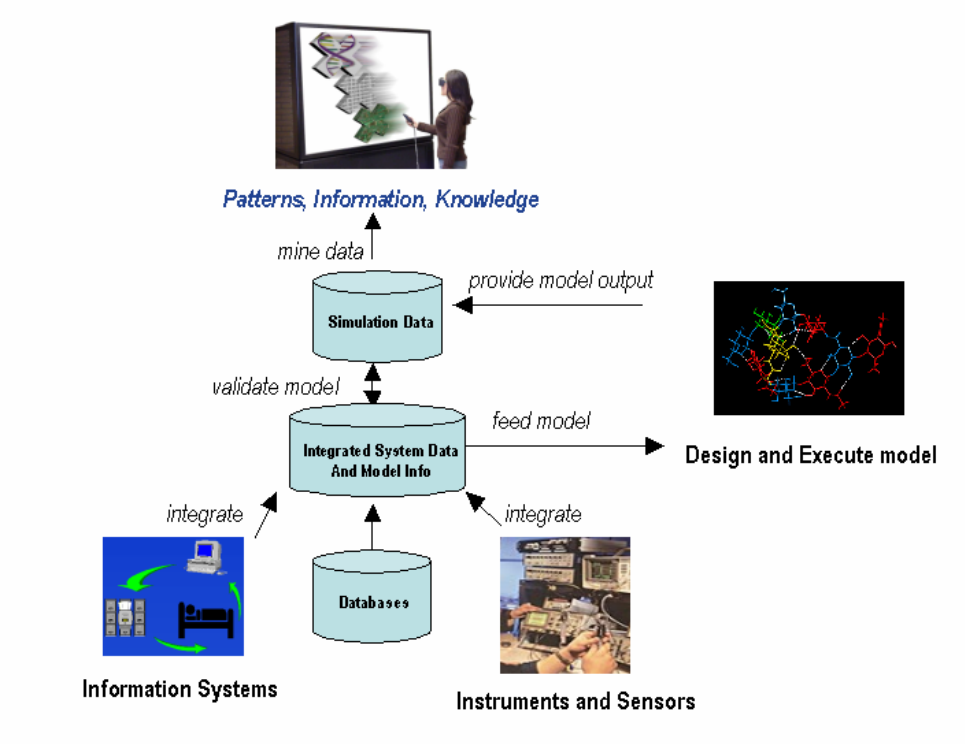


Figure 1 General architecture for conducting e-Science research: information systems integrate available data with data from specialized instruments and sensors into distributed repositories. Computational models are then executed using the integrated data, providing large quantities of model output data, which is mined and processed in order to extract knowledge. (Courtesy P.M.A. Sloot et al., IEEE computer 2006).

The new modeling approaches are increasingly based on *actual* and *detailed* data on the activity of individuals, their interactions and movement, as well as the spatial structure of the environment, transportation infrastructures, traffic networks, spreading of epidemics, 'omics' data, etc. Also, a huge amount of data, collected and meticulously catalogued, has become available for scientific analysis and study by the complex systems community (e.g. full details of road networks or air traffic). Regulatory networks capturing the mutual interactions between proteins in cells, networks which trace the activities and interactions of individuals, social patterns, transportation fluxes and population movements on a local and global scale have been analyzed and found to exhibit complex features encoded in large scale heterogeneity, self-organization and other properties typical of complex systems. Along with the advances in our understanding and characterization of systems' complexities, increased CPU power has led to the possibility of simulating multi-scale dynamical models consisting of thousands of coupled stochastic equations and allows agent-based approaches which include millions of individuals. Complex systems science is finally maturing and should now become data-driven and able to *predict* quantitatively the behavior of very large scale multi-level natural and artificial systems.

Despite such an integrated approach being in its infancy, the level of realism and detail achievable today allows us for the first time to ambitiously imagine the creation of computational infrastructures able to provide reliable, detailed and quantitatively accurate predictions for complex systems. These will be based on large, possibly distributed, data collection and multi-scale computational models as now used in weather forecasting. In other words computational approaches are now ready to

interface with the complex features of biological systems, infrastructure networks and social dynamics, entering the era in which they must become a major predictive tool. Such an approach will provide radically new ways of understanding the physical, biological, ecological, and social universe.

Nowadays complex systems science needs to bridge the gap between the individual and the collective: from genes to organisms to ecosystems, from atoms to materials to products, from notebooks to the Internet, from citizens to society. It cuts across all the disciplines. It is part of every discipline, see Figure 2.

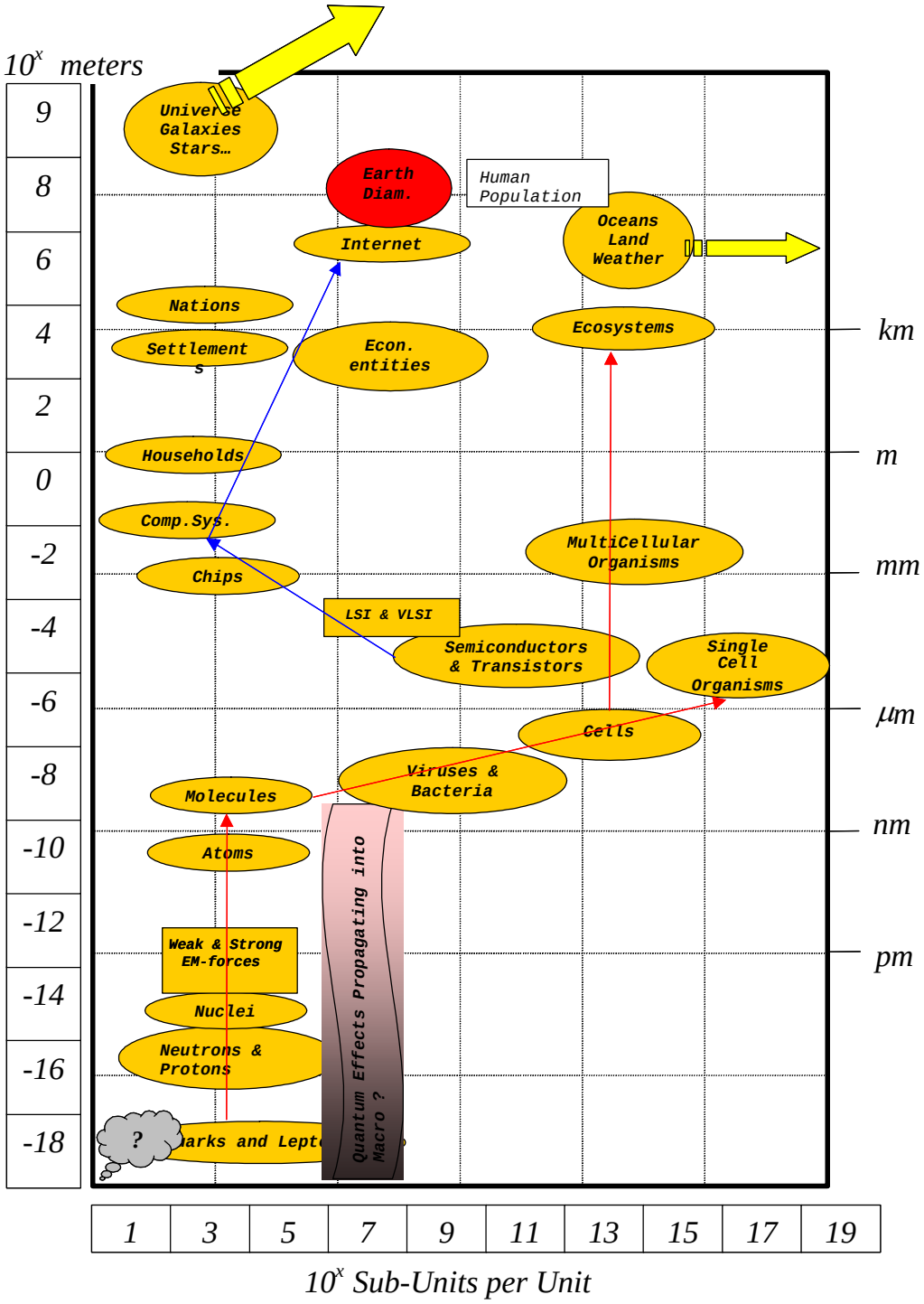


Figure 2: Scales in the Universe

Theme I: Scales in the Universe

Multi-scale modeling in biomedicine: from molecule to man

Impact on Science and Society
The radical increase of *in vivo* data on living systems enabled by new high-throughput experimental technologies and ICT creates completely new challenges. The enormous *size* of databases (petabytes, 10^{15}) will transform biomedical science. Heterogeneous (real-time) data from living organisms such as human, combined with new and more powerful ways of exploiting this wealth of data, will lead to major advances in our knowledge of the multi-level dynamics of living systems. This will greatly increase our capacity in design, planning, and management. For example, in health the new science of complex systems will revolutionize the medical treatment of diseases, and revolutionize the delivery of treatment [e.g. Sloot 2005, [18]]. Individual problems of individual people will be treated, creating the long-term Grand Challenge of 'Personalized Health'. This requires (i) huge distributed databases of every individual's genotype, phenotype, medical and general history, (ii) new ways of searching, communicating and processing this information, and (iii) new and more efficient ways organizing the delivery of food and health services to Europe's half billion inhabitants.
In this context, there is a strong drive for complex systems in all domains to generate new computational disciplines such as *computational neurosciences, computational biology, computational physiology, computational ecology, computational economics, computational sociology*, and so on. Necessarily these computational theories are 'local' to their domains and do not address the traversal questions of complex systems science. In particular they do not address the generality of reconstructing multi-level dynamics of systems from data collected at many levels.

Scientific opportunities
The wealth of data now available from many years of clinical, epidemiological research and (medical) informatics, advances in high-throughput genomics and bioinformatics, coupled with recent developments in computational modeling and simulation, provides an excellent position to take the next steps towards understanding the physiology of the human body across the relevant 10^9 (nm to m) range of spatial scales and 10^{15} (μs to human lifetime) range of temporal scales and to apply this understanding to the clinic (Figure 3; Ayache, 2005, [2], Hunter and Borg, 2003, [10]). Examples of multi-scale modeling are increasingly emerging (see for example, Davies, 2005, [5]; Iribe, 2006, [11]; Kelly 2006, [14]; Sloot, 2005 [18]). Human diseases affect structure-function relations at many levels (gene, protein, cell, tissue, organ, and organism) and a multi-scale modeling framework could provide a more rational basis for diagnosis and treatment than currently exists. Multi-scale modeling could potentially even result in applications for safer or healthier nutrition.

The challenge
The challenge is to develop mathematical and computational models of structure-function relations appropriate to each (limited) spatial and temporal domain and then to link the model at one scale to a more detailed description of structure and function on the adjacent levels. It is not obvious how to achieve continuity in the models and simulations across biological levels, each of which is normally governed by specific rules and assumptions. Bridging the gap between molecules and organ systems requires databases and knowledgebases of models at many spatial and temporal levels and software tools for authoring, visualizing and running models based on widely adopted modeling standards.

The data explosion
Biomedical and clinical research largely focuses on advancing our understanding about (patho)physiological processes in the human, which facilitates the further improvement

of diagnostics, treatment and drug discovery. Significant advances in molecular biology provided us with advanced experimental techniques to determine complete genomic sequences, measure gene, protein, and metabolite profiles, determine gene mutations, SNPs, protein interactions and epigenetic effects. Many of these techniques measure genome-wide scale and thereby provide a basis for 'systems biology' as the discipline that aims to understand the full dynamics of a cell. However, if we want to be able to understand the full physiology of the human as a basis to develop novel clinical applications we necessarily must go beyond the molecular level. Microscopic and imaging technologies such as microscopy, MRI and computer tomography are used to visualize the (sub)cellular and tissue level. *In situ* hybridization in combination with microscopy allows the determination of the spatial distribution of expressed proteins. Other medical techniques such as electrocardiography provide further information about organ structure and function. Finally, for patients (and healthy persons) a wealth of information about life style, disease history, genetics, plasma metabolite levels, clinical parameters and cell counts is generally stored in hospital databases and (electronic) patient files. Multi-scale modeling requires that we will be able to integrate and use this data.

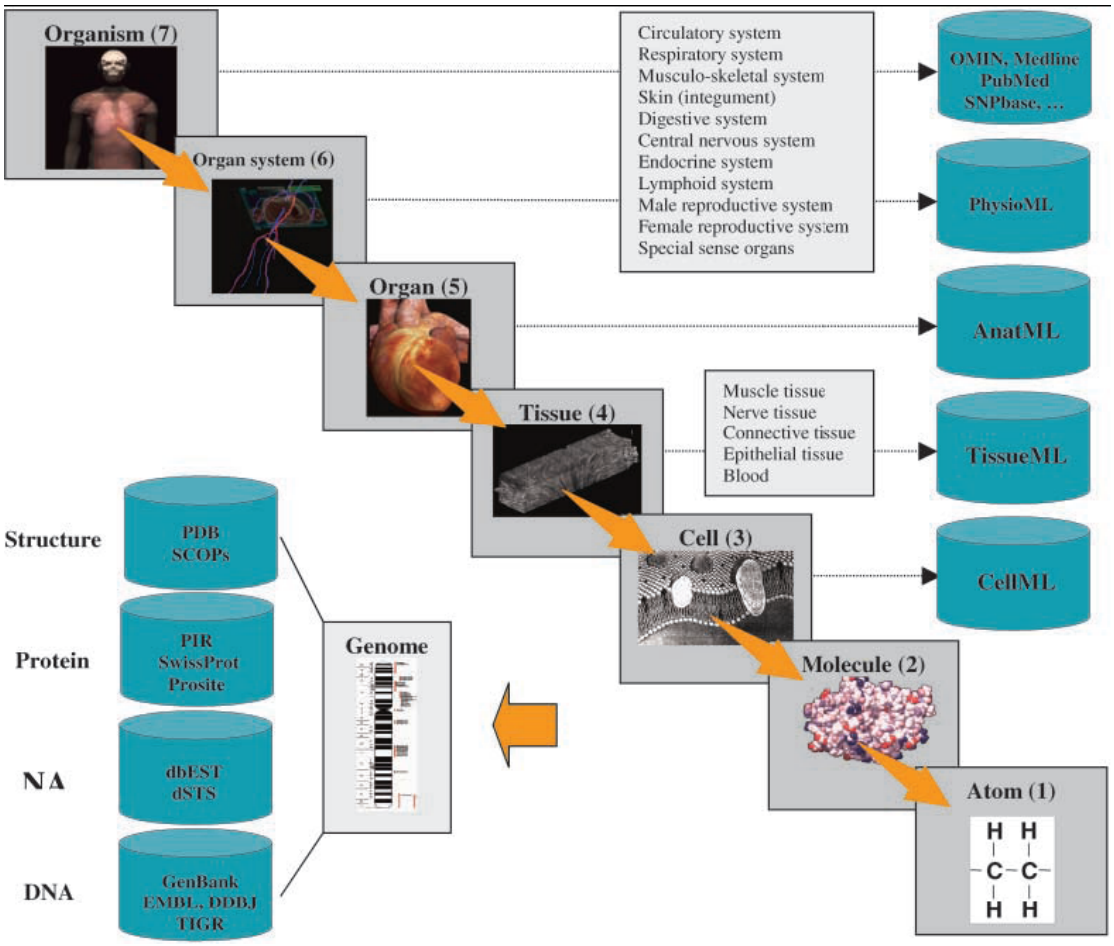


Figure 3. Accessing information at the various spatial scales. The markup languages (ML) ensure that models are encoded in a consistent form and allow simulation packages to import the models in a standard format. Databases of clinical information and databases of genomic and proteomic data are increasingly available (taken from Hunter P, Robbins P, Noble D (2002) The IUPS human physiome project, *Eur J Physiol*. **445**, 1–9)

Clinical applications

A solid validation and clinical evaluation of the '(patho)physiologic models' output versus actual (patho)physiological data is a necessary condition for the adoption of virtual human models by the medical community. The most important aspect of all is availability of high quality biomedical data for parameterization, validation and verification of the

models. Consequently, multi-scale studies of factors affecting the health and illness will require the involvement of (systems) epidemiology and biostatistics.

Organizing information and knowledge
Given the wealth of (qualitative) knowledge and data that is already available and necessary for multi-scale modeling, there is an urgent need to develop knowledgebases that provide 'static multi-scale models' of biological networks, cells and organs and their interaction. This requires further re-organization and integration of this information and novel approaches for representation and presentation of this information. Constraints and conditions for mathematical modeling and simulation can be derived from these knowledgebases and will provide the context for further analysis and interpretation of results. This implies a necessity to create a common language (ontologies) able to accurately describe all the complex structures and functions encountered in understanding and simulating living humans.

Computational approaches towards multi-scale modeling
The quantifiable framework that links molecular and cellular events with physiological function must be based on mathematical modeling as the language for describing physical and chemical processes. These mathematical models can subsequently be used to understand the dynamics of the systems and the interaction between the different temporal and spatial scales.
To overcome the different spatial and time scales mathematical (linked) models at the level of the organ, tissue and cell need to be developed. The review paper of Hunter and Borg (2003) [10], gives some requirements for these models. Equally important is the 'static' modeling of biological systems as is common in bioinformatics. One of the important areas of bioinformatics is the reconstruction and completion of biological networks (metabolic, signaling) from (public) experimental data. The development of accurate static biological models is a crucial prerequisite for further mathematical modeling and simulation.

Research Issues
A range of research issues were already identified in the VPH white paper (Ayache et al., 2005) and span from better use of existing data, methods and tools to the development of new methods, standards, libraries and tools for future research.

- *Libraries, databanks and data collections* are considered essential resources to be built. Further to gene, protein, tissue and medical records databanks, there is an urgent need for biological/biomedical-knowledgebases showing levels of evidence and range of observed variation and reference sources. This needs proper ontologies, fill-in and quality control procedures.

- The *research infrastructures* needed, will require Grid-enabled algorithms capable to exploit distributed (computing and storage) resources in specific Dutch and International networks, with a specific difficulty associated to the federation of databases that are owned by hospitals and clinical centres. Several complementary Dutch initiatives (e.g. BIG GRID [4], Parelsnoer, CTMM) are addressing this issue.

- Several *tools needed for modeling and simulation* have been identified. They are either basic tools needed to process basic issues underlying multi-scale modeling or related to specific modeling and simulation issues. Examples of the first category are mathematical models, data fusion tools, informatics for pathway elucidation, knowledgebases, cross-level ontologies, navigation tools, image computing techniques and how to deal with missing data or data uncertainty. In the second category examples such as model learning and hypothesis generation from integrated medical and genetic data, predictive models of disease and drug effects under different conditions of genetic inheritance and environmental interaction,

models of human organs including new paradigms to approach multi-level spatio-temporal changes such as, for instance multi-agent based modeling.

- *Standards, interoperability, proofs of concept, as well as pilot and deployment projects* are identified as major challenges. Benchmark datasets, standards for validated models and methods for performance evaluation will be needed if models are to be used by others than their designers. But also the method for their use should be validated and easy to apply. Proofs of concept, i.e. *in silico* developments dealing with practical issues, are needed to show that the aim is achieved or at least achievable in well chosen and meaningful cases. Many developments are in the transition between research use and clinical use. Therefore pilot studies performed by the technology/software developers in close collaboration with biologists and clinicians interested in leading edge developments pave the way for research-to-clinical use transition. The biggest related challenge seems to be developing clinically-friendly user interfaces in close collaboration with end-users.

Resources required
The most important requirement is the availability of data, models, expertise, etc, about a single biological system (e.g., cancer, HIV, etc) to truly engage in multi-scale modeling.

Metrics of success
The following metrics of success are distinguished in the multi-scale modeling in biomedicine: (a) establishment of a grid-based collaborative environment facilitating multi-scale modeling in biomedicine for specific research/clinical questions, (b) Effective collaboration between scientists from different disciplines using the grid-infrastructure, (c) the establishment of computational models for several time/length scales and the integration thereof, (d) one or more examples of e.g., improved diagnosis, (e) a national repository of data and computational models.

From particle to population: scientific computing in a distributed data environment

Impact on science and society: discrete problems
Applications ranging from protein folding studies based on particle simulations to the study of DNA microarray data involve huge quantities of data, often with complex relations between them.

In recent years, the multi-level method has been shown to be a very effective way of organising data in the context of *combinatorial scientific computing*, which deals with discrete problems occurring in scientific computing. This method reduces the size of the data space by first combining similar objects in a *coarsening phase*, repeating this until the data space is sufficiently small, then performing the required operations, such as partitioning or clustering the combined objects, and finally refining the result at successive levels during an *uncoarsening phase*.

Complex relations between the data are often captured by graphs, where nodes may represent for instance documents in an annotated database, or particles in a molecular dynamics simulation, and edges represent pairwise connections between them. Recently, hypergraphs have emerged as an even more expressive concept, which allows more than two nodes to be related, thus capturing more complex sets of relations. A hypergraph consists of a set of nodes (vertices), and a set of hyperedges, which are subsets of the set of vertices. If all hyperedges contain two vertices, the hypergraph reduces to a graph. Effective data partitioning methods have been developed based on hypergraphs, for instance to reduce the communication volume in parallel computations with iterative solution methods on supercomputers for large sparse matrices

(Vastenhouw and Bisseling, 2005 [22]). Such computations form the heart of many applications.

Impact on science and society: continuous problems
Virtually every branch of technology is relying on simulations where physical models leading to mathematical formulations, usually in the form of partial differential equations, are numerically solved. Typically, such problems have a variety of length and time scales, necessitating methods to cope with such properties. Examples are turbulence where a cascade of scales is present. Depending on the application, resolution for the larger scales only may suffice or rather the finer scales need to be resolved (Direct Numerical Simulation or DNS). To this end multi-scale and multirate methods need to be developed further, as in the various areas of the application domain different resolution is needed at the same time.

Another typical problem is given by the scales present in studying materials. On a microscale, atomic or molecular interactions can be described by multiparticle systems processes, typically leading to very large systems of ordinary differential equations (ODEs). For these, MD methods such as (quasi) Monte Carlo methods are needed, which call for parallel computations and efficient data handling. The biggest challenge is to link insights on the micro level to meso level constitutive properties like viscosity or why material is brittle or flexible (cf. glass). Here upscaling requires proper insight in representative volumes and most likely a probabilistic approach. Upscaling further, one encounters the macroscale problems, where this material is used in larger constructions. Studying such questions requires novel numerical methods as well as proper use of larger computational infrastructures. Multi-scale problems are also a central theme in the recently established Centre of Excellence (3-TU) "Multi-scale Phenomena". Mathematical and numerical modeling for a broad range of applications in industry is carried out at the centre for Analysis, Scientific computing an Applications (CASA-TU/e).

Scientific opportunities
Data clustering and partitioning are core operations that would benefit from recent advances in multi-level and hypergraph methods. They are generic and would be useful in a wide variety of applications, thus providing algorithm and software reuse, and leverage of scientific computing methods across different disciplines.
Grid computing is a natural extension of parallel computing that has emerged in the past decade, providing in principle tremendous computing power across the internet. It is a golden opportunity for expanding the size of large-scale simulations.

Research issues
A gap exists between the tightly-coupled massively parallel supercomputers such as those located at SARA and departmental PC clusters on the one hand and the more loosely-coupled grid computers working together on the other hand. The challenge here is to bridge this gap by generalising methods from the parallel computer case, to the situation with different processor speeds, varying communication loads and so on. More flexible data partioning is needed that can adapt itself to such situations. Such partitioners should also be able to tell where data must be stored (or relocated) for better retrieval.
A major effort will have to go into reducing the complexity of the problems. So called model order reduction methods (MOR) open up possibilities of tackling otherwise far too complicated problems. They try to limit the very large number of degrees of freedom in a problem to a much smaller set, related to the most relevant parts of the system which are involved in a process. Typically the number of elements on a chip (which may be of the order of a billion) that may play a role in finding out about a certain functionality can thus be reduced to a few (say 50). This then opens up the possibility to perform simulations in reasonable time for the latter.

Resources required
Needed is an investment in the development of algorithms, software, and testing in different application fields, gradually building up from the easier to the harder applications. Robust and efficient open-source software with proper documentation should be seen as a form of publication and should be rewarded as such. Large-scale computing facilities, both cluster-type and supercomputer-type, will be required to test algorithm efficiency and scalability over a wide range of applications.

Metrics of success
Success of developed algorithms and software should be measured by demonstration in at least two different types of applications, from different fields. A bonus would be breakthroughs achieved through simulations using this software.
The role of mathematics through scientific computing, probably even more impressive than Moore's law, cannot be overrated (see Fig. 4). It provides for much cross fertilisation, a role which becomes even more important when multi-physics is involved. Transfer of methodologies developed in one application area to another is very natural.

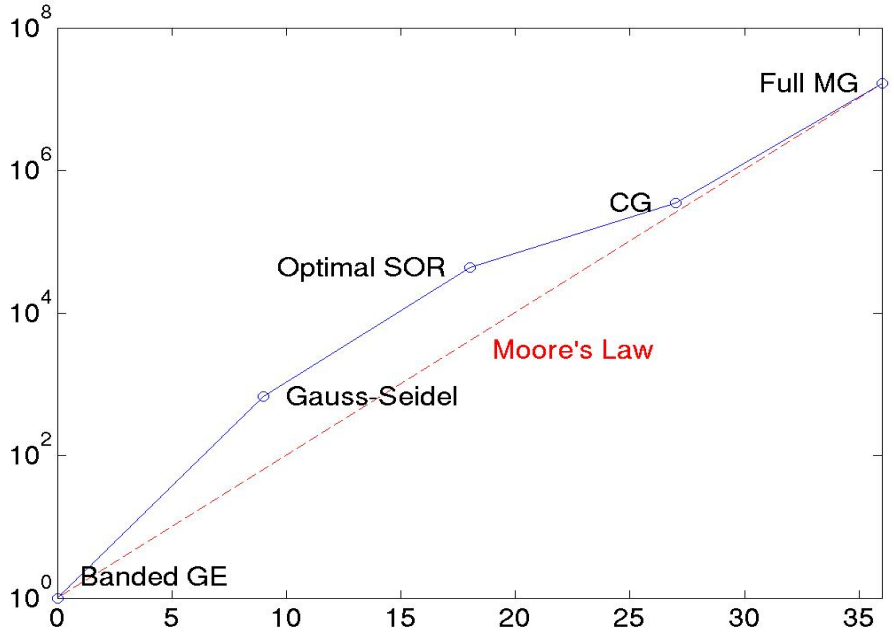


Fig. 4 Relative speed-up versus number of years: Algorithmic improvement for iterative solution methods for linear systems compared to hardware improvements. Source: Scales report, Vol. 1, Chapter 5 *Enabling mathematics and computer science tools [17]*.

Multi-scale Molecular Modeling

Impact on Science and Society

Multi-scale Molecular Modeling has evolved into a core discipline in theoretical chemistry, materials science and, increasingly, systems biology. For example, only recently the VU, UvA and AMOLF have joined to form the Amsterdam Centre for Multi-scale Modeling (ACMM), which has the express purpose to consolidate the Dutch international position in this field, expand the training of young researchers and strengthen the link between computational and experimental academic and industrial research. The Centre already plays a prominent international (EU funded) role in teaching multi-scale modeling techniques. It will also provide the environment in which experimental or industrial groups can obtain support for their modeling activities. At the ACMM, computer

programs and hardware will be made available for these researchers and they can obtain state-of-the-art advice from the staff members. Postdoctoral researchers in the Centre will spend part of their time supporting these researchers to carry out their research activities at the ACMM.
In quantum chemistry, the ADF package developed by Baerends and collaborators has evolved to become the international standard. It is commercialized by Scientific Computing & Modelling NV (SCM), a company spun off from the Baerends group in 1995. The book Understanding Molecular Simulation by Frenkel and Smit is one of the most widely used textbooks in the field of Molecular Dynamics and Monte Carlo modeling.

Opportunities
Internationally, the breadth of scope of the ACMM modeling activities is unique: the Quantum Chemistry group at the Free University specializes in the development of novel quantum mechanical tools (in particular, density-functional theory) to predict the properties of small numbers of atoms and molecules. The group at the University of Amsterdam specializes in the development and Monte Carlo and Molecular Dynamics simulation techniques to study systems containing many thousands of particles and on coarse-grained modeling of systems containing billions of particles. Finally, the group at AMOLF focuses on the development of novel computational tools to model growth and regulation in living cells. Strong contacts exist between the groups and it is this synergy that enables the expansion of research into emerging directions, most notable computational systems biology, computational catalysis and computational nano-science.

Research Issues

An attractive approach to multi-scale modeling, is the frozen density embedding (FDE) scheme that is based on an exact subsystem formulation of Density Functional Theory (DFT). The group at the VU (Visscher) is a driving force behind the development of this method, that is now mainly applied in predicting solvent effects on molecular properties. The research effort focuses on the implementation of energy gradients that will make the method a viable tool for ab initio dynamics on solvated complexes. Complementary to the FDE approach, the QUILD (QUantum-regions Interconnected by Local Descriptions) approach is being developed by the VU group (Bickelhaupt). The QUILD approach primarily aims at describing diverse physical and chemical phenomena (H bonding, stacking, chemical reactions) that occur in a large biochemical model system (DNA-template/DNA- or RNA-primer/enzyme complex) that serves to elucidate the DNA replication mechanism on the molecular level. Such insight is of direct importance for our understanding of (and our ability to repair defects in) gene expression and cell division (e.g., cancer research).

At AMOLF, the center of gravity of the computational research moves towards coarse-grained simulations of organization and transport of biomolecules and materials containing bio-molecular building blocks. One key line of research is the study of novel DNA-coated colloids. This work is carried out in synchrony with a parallel experimental program. In addition, the AMOLF research focuses on methods to gain a better understanding of the role of charge on the crystallization of proteins and (nano) colloidal salts. A third line of research focuses on “signal processing” in biology. This research will incorporate both stochastic modeling of gene regulatory networks and the coarse-grained description of individual signaling and translocation events. This involves the study of substrate-induced conformational changes in simple models for proteins. Apart from enhancing our understanding of biological regulatory systems, this research should help us to understand how to ‘design’ artificial binding sites with ultra high specificity.

This work is closely related to the research at the UvA on ordering of membrane embedded biomolecules; protein signal transduction, protein self-assembly and intercellular transport. These processes are intrinsically multi-scale and will be studied using a hierarchical approach from the quantum to the mesoscopic level. A second key

line of research at the UvA is in computational materials science, where the focus in the future will be on transport in nano-porous compounds and the structure and designing of carbonic compounds under extreme conditions. In this field, strong collaborations exist with AMOLF (Lattice Boltzmann modeling of nano-scale flow and study of carbon nucleation).

Resources required
Needed is access to advanced capacity computer systems with large memory and storage.

Theme II: The national e-science fabric

Introduction

The second theme studies the national e-Science fabric. Here the goal is to develop a computational grid infrastructure for collaborative scientific research of complex phenomena and processes on a wide range of scales. This is approached from a number of different perspectives.
First, we use a major application research area with extremely demanding needs for computing and data management: astronomy and astrophysics. The (international) development towards virtual observatories, the hugely compute- and data-intensive simulations, and the coupling between these simulations and real observations taken by modern fast data collecting instruments make this field an obvious candidate for e-Science.

Although grid software has been developed and many experiments have been performed, the software engineering aspects of grid software received much less attention, often resulting in software that is not reliable enough yet for large scale deployment. Therefore, special attention is paid to methods and techniques for achieving classical software engineering quality attributes like availability, modifiability, performance, security, testability and usability and applying them to grid applications. Applications like astronomy have challenging demands on the underlying middleware. Foremost, the middleware should be robust and subject to strong software engineering quality metrics. Many grid middleware systems lack a clear architecture for the application, which complicates programming and interfacing. Also, the sheer size of the computations requires a level of scalability of grid middleware that simply does not exist yet. Likewise, fault tolerance becomes a major concern at this extreme level of parallelism.
In order to understand complex multi-scale phenomena in astronomy, biology and other sciences, visualization is critical, not only to aid in understanding the individual scales, but increasingly, also to (literally) show the big picture by connecting the spatial and temporal scales in a few images or an animation. Since the data sets are often extremely large, distributed and collaborative visualization using computational grids becomes increasingly important. For quality control and exploration of parameter spaces of complex observational instruments (telescopes, scanners, etc.), visualization proves to be useful as well. Visualization research has many fundamental aspects, but it also is a discipline that needs close collaboration with scientific applications.
A further recent technological advancement is the use of Wireless Sensor Networks for remote monitoring applications. Such networks allow detailed measurements in natural environments from local to global scales, which can be coupled to computational models.

Some of the *infrastructural* aspects mentioned here, like the required computational resources and data storage facilities, are available or under construction, as foreseen in NCF's activities and in the granted BIG GRID proposal for setting up a Netherlands Science Grid [BIG Grid, [4]].

Distributed systems in astronomy and astrophysics: 100-Teraflop simulations and Petabyte observations

Astronomy has always been a highly data-intensive science with strong computational demands, both for the processing and analysis of observational data and for the simulation of complex astrophysical phenomena. Astronomers are used to working in widely distributed research communities who collaborate closely in order to reach common scientific objectives. Nationally, astronomers at universities and other research centers collaborate within the top research schools NOVA, ASTRON, JIVE and SRON; internationally, the data flood is channeled with the help of the European Virtual Observatory (Euro-VO). The Euro-VO sets international standards for data formats and

manages a central registry for archives (hosted by the European Space Agency) which enable users to query multiple astronomy databases all over the world with a single query. Using this infrastructure a wealth of web-services and tools has been created, often by National Virtual Observatory Organizations.

This year (2006), the European Virtual Observatory Data Center Alliance (Euro-VO-DCA- with participating countries F, G, I, Sp, UK, NL, ESA and ESO, with NOVA representing the NL astronomical community) has started a new initiative to stimulate and enhance VO technology, to make data archives VO compatible, to connect the VO network to the GRID and, for the first time, to include the results of huge numerical simulations into this environment. For example, the Millennium simulation of the universe with 10 billion particles (350.000 CPU hours) has been prepared for virtual observing by the German Virtual Observatory and the Euro-VO. Web services designed for virtually observing the simulated universe can be connected to services that query databases of true observations, facilitating direct comparisons between theory and observations.

The next step will be the integration of the *production (supported by a compute grid*) of the archival qualified observational and simulated data and the *access to the results (supported by an info grid)* into a single VO research environment, thus creating information systems which provide a true VO research environment This will allow critical scientists to re-process archived results, starting from the raw data, and to publish new results into the VO. For both observations and numerical simulations, the ability to reprocess and reanalyze data is essential because it allows scientists to ask new questions and to apply new data analysis methods. This also fits naturally in the data processing in the radio regime, where the Dutch have traditionally played an international leading role.

An EU wide pioneer project, AstroWise, led by NOVA-OmegaCEN, has developed such a system for the next generation of wide field optical imagers, which will produce orders of magnitude more data than present surveys, and numerical experiments.

The move towards integrating VO and processing oriented services is also an important issue for radio observatories. In the next years new telescope initiatives, LOFAR, eVLBI and eventually SKA will produce data sets that no longer fit the processing model where the scientists process data locally, finding the optimal processing parameters by trial and error. Some of the algorithm research required for this has been carried out in ALBUS, under the EC funded RadioNet project and has been led from JIVE, Dwingeloo.
This envisioned integration of astronomical observations and numerical experiments is very ambitious and will be exemplary for other science projects dealing with truly distributed processing and with archiving of very large data sets.

Opportunities NL-VO

In this landscape, a new National Coordinated Initiative would be very well positioned to support the equivalent of a NL-Virtual Observatory, which could develop and operate a national astronomical research environment for the next generation of numerical simulations and Petabyte observations. Indeed, the flood of data delivered by upcoming astronomical facilities, such as the LOFAR radio telescope, the eVLBI network, the OmegaCAM wide field imager and the GAIA satellite, will require fundamentally different approaches at virtually all steps of the scientific discovery process, from the raw data taken at the observatories to the interpretation by the end user. All this can be classified under the discipline “Computational e-Science”.

Research issues

Below, we summarize some key astronomical experiments envisaged to be pooled in a National Coordinated Initiative.

LOFAR – "first light in the Universe"

The **LOFAR** radio telescope, which is currently being constructed in the Netherlands and is scheduled to be fully operational in 2008, will generate a raw data stream of 320 Gb/sec, sustained. Obviously, such data streams need to be reduced to a more manageable data rate in an unattended real-time manner. The output data rate will be some 10 TByte/day, which is still significantly larger than the amount currently being handled by the astronomical community. For LOFAR this will be handled by a dedicated central processing system, consisting of an 12,288-processor IBM BlueGene/L system in Groningen, together with a large Linux cluster for applying complex calibrations (ionospheric and instrumental models), which require feed-back from the total data volume ('sky model'). The feedback loop between calibration and scientific results demands complete integration of the compute and knowledge resources and is therefore typical of the objectives of this initiative. For example, the "Epoch of re-ionization" key project will collect ~1 Petabyte of data, which will have to be reprocessed over and over until all contaminating foregrounds are removed, leaving only the radiation emitted 13 billion years ago by intergalactic gas during the epoch of 'first light'.

eVLBI, beyond LOFAR and on the way to SKA

Beyond LOFAR, the eVLBI array of the future will produce many TBytes for single user experiments. Eventually the Square Kilometer Array will be constructed for which the data streams will even be manifold of the above. LOFAR has already adopted new user models, where it is assumed that most users will deal with calibrated images which will be disseminated through the VO. This is different from the current practice in which astronomers process data on local workstations with public legacy software. This will, however, not satisfy all cases. There will be scientists who want to explore the parameter space of the reduction process in order to reach the limits of what the data allows. For the telescopes of the future this will not be done in the user domain. The data should be accessed from the archive and CPU cycles will be found in the Grid. Establishing new, more robust data recipes for such large data processing jobs and making them available in a distributed fashion is an issue in modern radio astronomy.

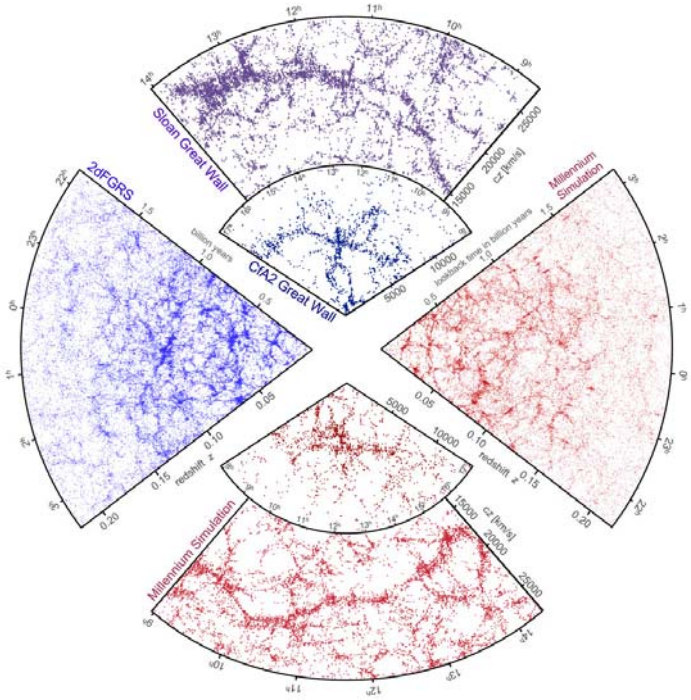


Figure 5: Large scale structure in the Universe. Each dot represents a galaxy: in blue-ish colours truely observed , in reddish the results of numerical simulations- courtesy Simon White –MPE

OWLS --"galaxy formation and evolution"
The **OWLS** (OverWhelmingly Large Simulations) project, which is led by NL astronomers, investigates one of the central topics in modern astronomy: the formation of galaxies and the evolution of the intergalactic medium. The complexity of the physics, which involves a large range of scales and many feedback processes, has made numerical simulations an invaluable tool in our quest for understanding. With the availability of huge computational resources (both NCF's facilities and the BlueGene/L system at the RuG), computational cosmology in the Netherlands is about to make a huge step forward. The OWLS project, which has just started, will provide the largest simulations of its kind and will generate ~50 TBytes of data using millions of cpu hours. The simulations will be analyzed by a large, international team of researchers and eventually they will be released to the whole astronomical community. This again requires scientists to have simultaneous access to significant computational and data storage resources.

OmegaCEN – KIDS "**dark matter, dark energy**"
The NL coordinated KIDS (KIlo-Degree Survey) project, which will start in 2007, aims to characterize the properties of mysterious dark matter and dark energy which dominate the dynamics of our universe. KIDS will accumulate a data volume that is an order of a magnitude larger than the Sloan Survey, which represents the current state-of-the-art. KIDS will be processed, analysed and disseminated by Nationally distributed research groups (Leiden, Groningen, Nijmegen), collaborating with partners in Bonn, Munich, Paris and Italy.
All these groups work together in a fully distributed compute and information grid supported by real-time synchronized databases. The KIDS project involves tens of millions of CCD read-outs, with 8 Mega pixels of raw data per read-out. For distributed processing and analysis, and for the integration of computational grids with distributed knowledge bases (databases), it is required to have many essential computational e-science innovations.

GAIA 10^9 stars -- 6 dim phase space
GAIA is an ESA mission, scheduled for launch in 2011, which will provide a stereoscopic census of the stars in our Galaxy through precise measurements of positions, velocities, and colors. The posterior on-ground data processing is a very large and highly complex task, linking all astrometric, photometric and radial velocity measurements into the so-called global iterative solution. In addition, radial velocities will be measured for all objects to 17^{th} magnitude, thus complementing the astrometry to provide full six-dimensional phase space information.
The processing activities will be structured around nine scientific units and six data processing centers. The NL GAIA community anticipates a key-role in the data mining and in the development of computational e-science technology for integrating compute grids and knowledge bases

Resources required
All these projects require large, integrated resources for computations and data storage, as well as a common infrastructure for accessing, sharing, and processing data. This requires powerful and intelligent distributed data servers and advanced information systems coupled to grids. Such systems do not yet exist and their development is a major goal for the initiative.

Innovative software will also need to be developed. First, there is a clear need for improved computational techniques to allow researches to analyze large data sets and numerical simulations. Second, accessing and sharing data among a group of (international) researchers requires advanced information systems, facilitating security and accessibility. Third, the visualization of massive, multi-dimensional data sets poses a formidable challenge.

A National coordinated initiative in Computational e-Science will be able to cover all these aspects. It will bring together researchers from various fields, enabling them to identify common problems and challenges and to share solutions, strategies, and resources.

From code fragments to software systems

Impact on Science and Society
A strong distributed computing infrastructure is essential for e-Science applications, since such applications typically access computers, databases, and instruments from many different places and are used by collaborating research teams at different locations. Extensive research has already been done on such distributed infrastructures, especially in the context of Grid computing. Grid computing tries to integrate geographically distributed resources in a transparent way. Grid software has been developed and many experiments have been performed. However, the software engineering aspects of Grid software received much less attention, often resulting in software that is not reliable enough yet for large scale deployment. This is illustrated by NSF's current strong support for the software engineering aspects in the Globus project.

A major step forward is possible when methods and techniques are created for achieving classical software engineering quality attributes like availability, modifiability, performance, security, testability and usability and applying them to grid applications. This can be achieved at the architectural level by providing mechanisms like transactions, secure computing mechanisms, load balancing and generic monitoring and measurement frameworks.
By creating a common architecture for e-science applications, the expertise from a wider range of scientific areas can be brought together. This requires defining an overall architecture as well as detailed workflows and Application Programming Interfaces (APIs). In this way, components from a wide range of disciplines can be developed that fit more seamlessly in the common architecture. The impact will be twofold. First, this new e-science infrastructure will enable new, and revolutionary, experiments and simulations that will cause breakthroughs in many areas. Second, the techniques used to achieve this infrastructure are of more general importance; they may have impact in all areas where computing agents interact. They will be additions to or replacements of approaches like distributed computing, web-services and service-oriented computing.
Spin-off to other disciplines that are running into the data avalanche can be expected. For example, the pilot astronomical information system AstroWise is now applied by the RuG artificial intelligence group to achieve production (Blue Gene) and distribution of pattern recognition of 50 Terabyte of scans of 100 years of handwritten text of the “Kabinet der Koningin’ of the National Archives.

Opportunities
There are many opportunities for software engineering to improve the state-of-the-art and deployment of grid technology. Foremost, most grid applications and middleware lack a clear global architecture. Often, middleware is designed as a ``bag of services'' and applications just select the services that are needed. The field of software architecture has made much progress during the previous years, and we think a cross-fertilization between the field of software architecture and grid technology could be highly beneficial. Examples are not only the application of state-of-the-art informal architectural design and documentation techniques but also the use of more formal approaches that enable the timely estimation of the quality attributes mentioned earlier. Another example are the transfer and adaptation of experimental middleware approaches (based, for instance, on blackboards, processes or channels) to the grid domain. In this way, also new analysis techniques for grid applications may become available.
A problem with deploying grid middleware is the fact that these systems are still

changed frequently, making them a moving target for applications. Globus, for example, changed from proprietary services (Globus Toolkit 2) to OGSA (GT 3) to web services (GT 4). As a result, the Application Programming Interfaces (APIs) are in a continuous flux, making it difficult to develop and maintain applications. To shield application programmers from the rapid changes, higher-level interfaces are needed that are more stable. A good example is SAGA (Simple API for Grid Applications) that is currently developed within OGF. SAGA tries to abstract services like job submission, file I/O, and monitoring.

As grids become more mature and are deployed on a larger scale, scalability itself becomes a critical factor. It is well known that grids have a much higher failure rate than traditional systems, due to their heterogeneous nature. As grids scale up, the likelihood of failures will increase further, making error detection and fault tolerance a key concern. Also, obtaining high performance (efficiency) on a very large scale system is far more difficult than on a small scale test bed. Many researchers believe that future grid systems should therefore be able to adapt, tune, and even heal themselves, because requiring human intervention will be undesirable or even infeasible. However, developing such ``self *'' algorithms and software still requires much new fundamental research, on the border of software engineering and grid technology. So, this area is a major opportunity for new research.

High-level workflow is crucial for the organization of grid applications. It determines the way in which computation and information flows through the grid infrastructure and is responsible for making the appropriate connections between distributed computations. However, it is hard to write, understand and debug these workflows. An additional problem is that global computations have to be implemented by local components and that it is hard to integrate the components at the global level with those at the local level.

There is therefore a strong need for component-based development of grid applications that enables this integration at both levels and supports a more structured development of high-quality applications.

Research issues

Based on the analysis of impact of and opportunities for applying software engineering techniques to the development of e-science applications we have identified two main promising research areas.

i) *Architecture*. The current status that grid applications consist of mostly incoherent services should be replaced by a situation that these applications are governed by a global architecture that clearly describes overall organization, purpose and qualities of each application. It is a challenge to identify the heuristics to arrive at such an architecture and to describe and validate it. This requires the development of methods for the architectural description of grid infrastructure and grid applications and the application of these methods to existing and new applications.

ii) *Component framework.* Component frameworks and software buses are needed that enable the integration of coarse-grain wide-area grid applications with fine-grain local area applications. Methods and techniques for rapid prototyping, debugging, testing and validation should form an integral part of such a framework. High-level workflow mechanisms are needed to integrate components into grid applications. These workflows should profit from basic mechanisms for error detection, error recovery, load balancing and the like that already are part of the overall architecture and that are provided by the component framework.

Many aspects exist that are orthogonal to the above two levels. Of these we discuss fault tolerance and scalability.

Fault tolerance: methods and techniques for error detection and error recovery are needed to improve the reliability of distributed applications. Since grid applications are wide-area, distributed, applications that cross many organizational boundaries the already difficult problem of achieving fault tolerance becomes even more complex. How is failure detected? Can a failed computation be moved to another computing site? How are the data of the failed computation handled? Can these problems be addressed in an autonomous (``self *'') way, without human intervention? These and other problems need innovative solutions.

Scalability: grids have a large potential for distributed, massively concurrent, computations. The truth is that this potential is not yet exploited at all. Today, most concurrent computations work on independent slices of data and are largely independent. Given the increasing bandwidths of communication networks, highly dependent distributed, concurrent, applications become feasible.

We think that the above issues should be studied from a coherent perspective by the various teams cooperating in this effort.

Resources Required

The above research issues will be investigated by a number of core teams, each team participating in the development of a common view on the e-science infrastructure but will specialize in a subset of the research issues mentioned above.
In addition, the following infrastructure will be needed:

- Access to relevant grid resources.
- Local computing clusters.

Since fault tolerance is one of the focus areas, it is important to have control over both local and global “failures” of grid services. Clearly, dedicated grid services have to be used for this.

Metrics of Success

The success of the research in this area will be measured by the following criteria:

- An operational common e-science infrastructure.
- Advancement of software engineering research regarding distributed applications.
- Application of these software engineering results in areas outside e-science.

Visualization of multi-scale processes

Impact on Science and Society

In scientific research, such as material science, astronomy, meteorology, biology and the medical sciences, the demand for visual interpretation tools has increased enormously now that computer simulations or computer-assisted measurement equipment produce amounts of data that are too large to be understood in numerical form or by simple graphs. This created the research area of Scientific Visualization or Data Visualization: its birth is usually associated to the ACM SIGGRAPH panel report *Visualization in Scientific Computing* from 1987. In this report, the goal of scientific visualization is defined as follows: "Visualization is a method of computing. It transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations. Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights. In many fields it is revolutionizing the way scientists do science."
Visualization is now routinely used in many cases, but it has not fulfilled its promises yet. Visualization is often hard to use in practice and does not support the users enough in the discovery process. However, if the integration of visualization with the concepts, processes, tools, and computing machinery of its users is improved, it can leverage many different fields by providing more insight in shorter time. Also, in view of the still increasing data sizes, visual representations will increasingly become the most practical

and important way to interpret the data and spot interesting patterns.

Opportunities

The big change since the report of 1987 is the *information big bang*, which is produced primarily in digital form. Understanding and using this deluge of information is regarded as the biggest challenge for the 21st century, and visualization is expected to play an important role in this. A recent report [Johnson et al., 2006, [13]]of the U.S. National Institute of Health and the National Science Foundation describes the challenges of the research area. On the one hand, many visualization algorithms are now standard available in commercial workstations and medical equipment, and users can handle these applications without support from visualization researchers. But many visualization applications fail because they are used wrongly or not at all. In short, the gap between domain experts and visualization has to be narrowed.

Besides these issues, also the visual representation of data itself continues to be a challenge. The data we have to deal with are large, time-dependent, uncertain, heterogeneous, multi-scale, and multi-variate. Multi-scale (or multi-resolution) approaches in visualization arise for two different reasons. First, because of the large sizes of (3-D) data sets, it is convenient to decompose data on various levels of detail, which can then be incrementally visualized (and transmitted, say via internet). For this purpose multi-resolution methods are being developed, based on techniques such as wavelets or pyramids well known in image processing. Such techniques allow the data to be described with less numbers, making these methods very useful for data compression.

Second, many natural phenomena involve a large spectrum of spatial and temporal scales. Astronomers study spatial scales that range from planets to galaxies; biologists study life processes from molecule to cell; brain scientists try to understand the brain from nerve cell to brain region. In all these cases, visualization can be of use, not only to aid in understanding the individual scales, but increasingly, also to (literally) show the big picture by connecting the spatial and temporal scales in a few images or an animation.

It becomes increasingly clear that in order to understand complex phenomena, the traditional data visualization techniques have to be complemented by methods from Information Visualization and Visual Analytics [Thomas et al., 2005, [21]]. The latter combines aspects of scientific visualization, human-computer interaction, data mining, pattern recognition, searching in large and/or distributed databases and internet technology to visualize abstract information.

Also, in order to move towards a public activity using shared data repositories, computational grids and distributed collaboration, a greater level of formalization is required, and meaning becomes a shared responsibility, for instance in the form of a shared ontology of the visualization process [Geroimenko et al., 2003, [8]].

Visualization of complex phenomena on multiple scales provides a challenge where we expect ample opportunity for addressing the issues raised recently in the visualization community. It is a problem with enormous potential, where many of the existing techniques fail. Contact with researchers from various application areas is essential. Success can only be achieved if scientific, economic, cognitive and social aspects are all taken into account.

Research issues

Based on the preceding analysis, a number of promising research issues can be identified. On a coarse scale, two approaches can be pursued: applied and foundational research. Concerning applied research, each domain (chemistry, astronomy, biology, medical sciences, etc.) has its own challenges and peculiarities. In close cooperation with domain experts optimal visualization solutions can be developed, where analysis of user needs, tight integration with their tools, insightful representations, intuitive user interaction and real-time parameter control are key issues, followed by thorough evaluation and validation in practice.

Generalization on the results, grounded in disciplines like computer science,

mathematics, and psychology, statistics, should lead to unified models of visualization in order to understand key success factors; methods and approaches to develop visualizations more effectively and efficiently; as well as frameworks and tool-kits to recycle results.
We expect fruitful cross-fertilization of the Computational e-Science program with two related NWO-funded research programs: VIEW (Visual Interactive and Effective Worlds), which focuses on generic visualization techniques, and STARE (STAR E-Science), which stimulates research on the interface between astronomy and computer science.

Resources Required
The following infrastructure will be needed:

- Access to relevant (grid) resources for collaborative visualization.
- Local computing and visualization facilities.

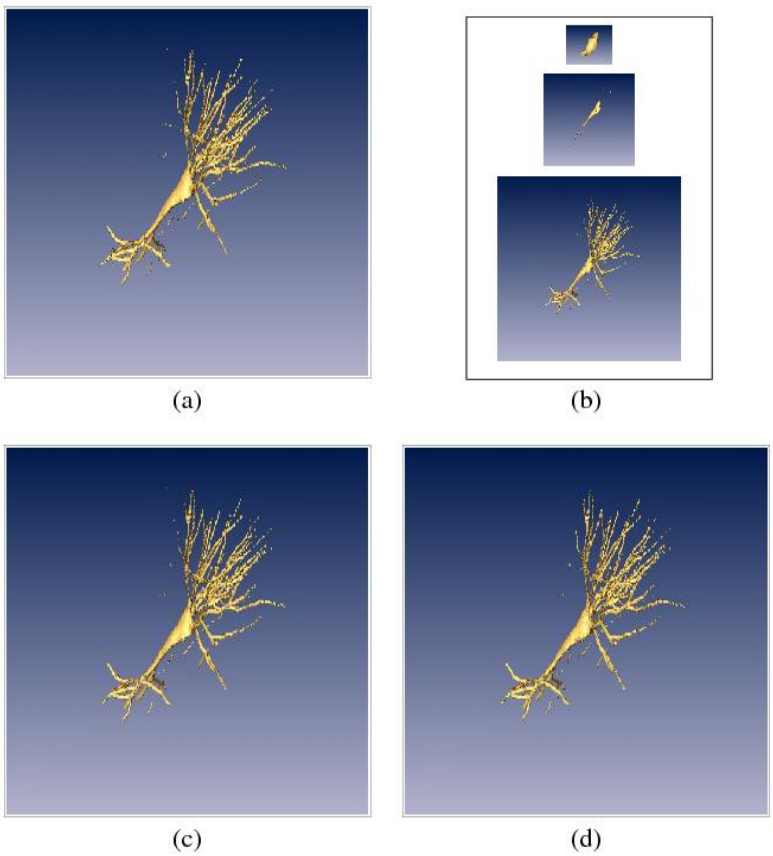


Figure 6: Multi-scale reconstruction of volume data improves interactivity due to data compression, here illustrated by visualisation of neuron volume data. First a multi-scale pyramid is constructed, by filtering the input volume and reducing the data size for each successive pyramid level. Then, during interactive visualisation a judicious selection of data from the pyramid is made so as to minimize memory usage while still satifying a given mean absolute error tolerance (MAE). A very fast reconstruction at full accuracy can be made whenever the user requires this. (a) Input data. (b) Three-level multi-scale pyramid. (c) Reconstruction with 80% data reduction, MAE=0.2. (d) Reconstruction with 95% data reduction, MAE=0.35. See: J. B. T. M. Roerdink, 2005 [16].

Metrics of Success
The success of the research in this area will be measured by the following criteria:

- Increased insight in the effectiveness of visualization, such that new applications can be developed in shorter time, preferably by end-users;
- Effective methods to visualize large, time-dependent, uncertain, heterogeneous, multi-scale, and multivariate data-sets;
- A number of visualization tools that are routinely used by domain experts.

From Sensor Networks to Decision Support

Impact on Science and Society
The ongoing advances in digital circuitry lead to ever smaller, ever cheaper, computer systems. This trend is being exploited by Wireless Sensor Networks (WSN) to combine sensing, local processing, and communication in individual sensor nodes, which can be deployed at large scale through a self-organizing network demonstrating collective capabilities currently unaffordable. Many potential applications have been envisioned (e.g., habitat monitoring, disaster management, object tracking, precision agriculture), mostly in the remote monitoring domain.

From a scientific perspective, WSN allows detailed measurements in natural environments on a scale and spatial resolution never seen before. For example, standard agricultural practice is based on a weather station about 20 km away; with WSN technology it becomes feasible to monitor the micro climate *inside* a field. This will eventually reduce costs and increase yield allowing society to make better use of the limited natural resources.

Opportunities
An interesting aspect of WSN is that data can now be fed in real-time into simulation and prediction systems. This allows decision support systems that are based on forecasting scenarios to update their expectations whenever new data becomes available. This novel opportunity of getting real-time data into the prediction loop holds a great promise for improving decision making processes. For example, a dense network of sensors measuring the local weather could be coupled to a model predicting the national weather.

Research Issues
Although the research field of WSN is rapidly maturing, the experience with pilot deployments is that there is a strong need for improving the design-flow and tools for developing and maintaining WSN applications. Currently, it takes highly skilled people a lot of time to hand code applications, and keeping a sensor network up and running is expensive because of software failing to handle unforeseen conditions. Thus, research into dependability and fault tolerance for resource-scarce devices is of great importance.

Bringing WSN technology into the basic loop of decision support systems will require research into a number of basic issues:

- simulation algorithms have to be adapted for computing with real-time data
- algorithms have to be made robust to handling imperfect data due to node and link failures within sensor networks.
- To reduce the raw data stream part of the simulations need to be performed locally at individual sensor nodes.

Resources Required
To demonstrate the feasibility of the coupling of sensor networks and decision support systems a prototype system must be developed. This should include a sizeable number (1000+) of sensor nodes linked over a wired backbone to a central computer running the main simulation.

Metrics of Success
The research will be considered successful when it brings the robustness of sensor networks onto a higher level, succeeds in adapting simulators to handle data in real-time

References

1. Afsarmanesh H., R.G. Belleman; A.S.Z. Belloum; A. Benabdelkader; J.F.J. van den Brand; G.B. Eijkel; A. Frenkel; C. Garita; D.L. Groep; R.M.A. Heeren; Z.W. Hendrikse; L.O. Hertzberger; J.A. Kaandorp; E.C. Kaletas; V. Korkhov; C.T.A.M. de Laat; P.M.A. Sloot; D. Vasunin; A. Visser and H.H. Yakali: *VLAM-G: A Grid-based virtual laboratory*, Scientific Programming, (Special issue on Grid Computing) vol. 10, nr 2 pp. 173-181. (R.H. Perrott and B.K. Szymanski, editors), IOS Press, 2002. ISSN 1058-9244.
2. Ayache N, Boissel J-P, Brunak S, Clapworthy G, Fingberg J, Lonsdale G, Frangi A, Deco G, Hunter P, Nielsen P, Halstead M, Hose R, Magnin I, Martin-Sanchez F, Sloot P, Kaandorp J, Hoekstra A, Van Sint Jan S, Viceconti M (2005) Towards Virtual Physiological Human: Multi-level modelling and simulation of the human anatomy and physiology. White paper, edited by Norager S, Lakovidis I, Carbrera M and Ozcivelek. http://ec.europa.eu/information_society/activities/health/docs/events/barcelona2005/ec-vph-white-paper2005nov.pdf
3. Barabasi A., "Taming Complexity", *Nature Physics*, Vol 1, November 2005, pp 68-70.
4. BIG Grid, the Dutch e-Science Grid, 13 October 2005, NCF, NBIC and NIKHEF.
5. Davies PF, Spaan JA, Krams R. (2005) Shear stress biology of the endothelium. *Ann Biomed Eng*. 33(12), 1714-8.
6. Finkelstein A., J. Hetherington, L. Li, O. Margoninski, P. Saffrey, R. Seymour and A. Warner, "Computational Challenges of System Biology", *Computer*, Vol. 37, No. 5, 2004, pp. 26-33.
7. Foster I., C. Kesselman and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations", *International Journal of High Performance Computing Applications*, Vol. 15, 2001, pp. 200-222.
8. Geroimenko V. and C. Chen, Visualizing the semantic web : XML-based internet and information visualization, Springer, 2003
9. Hey A.J.G. and A.E. Trefethen, "The Data Deluge: An e-Science Perspective", *Grid Computing - Making the Global Infrastructure a Reality*, chapter 36, 2003, pp. 809-824.
10. Hunter PJ, Borg TK (2003) Integration from proteins to organs: the Physiome Project.*Nat Rev Mol Cell Biol*., 4(3), 237-43.
11. Iribe G, Kohl P, Noble D. (2006) Modulatory effect of calmodulin-dependent kinase II (CaMKII) on sarcoplasmic reticulum Ca2+ handling and interval-force relations: a modelling study. *Philos Transact A Math Phys Eng Sci*. 364(1842), 1107-33.
12. i-science cluster: http://www.acts-nwo.nl/nwohome.nsf/pages/NWOP_65YM49
13. Johnson C.R., R. Moorehead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, (Eds.): NIH-NSF Visualization Research Challenges Report; IEEE Press, ISBN 0-7695-2733-7, 2006. http://tab.computer.org/vgtc/vrc/index.html
14. Kelly D, Mackenzie L, Hunter P, Smaill B, Saint DA. (2006) Gene expression of stretch-activated channels and mechanoelectric feedback in the heart. *Clin Exp Pharmacol Physiol*. 33(7), 642-8.
15. LMS-Virtual Laboratory: http://www.lmsintl.com/virtuallab
16. Roerdink J.B.T.M., "Morphological Pyramids in Multiresolution MIP Rendering of Large Volume Data: Survey and New Results", J. Math. Imag. Vision 2005, vol. 22, pp. 143-157.
17. Scales: A Science Case for Large Scale Simulation: http://www.pnl.gov/scales/
18. Sloot P.M.A., Boukhanovsky AV, Keulen W, Tirado-Ramos A, Boucher CA. (2005), A Grid-based HIV expert system. *J Clin Monit Comput*., 19(4-5), 263-78.
19. Sloot P.M.A., I. Altintas, M.T. Bubak, A. Tirado-Ramos, C. Boucher. 'From molecule to man: the system science of decision support in individualized e-Health', IEEE computer (cover feature), November 2006, pp 40-46.
20. The Netherlands Organization for Scientific Research (NWO): www.nwo.nl
21. Thomas J.J. and K.A. Cook (eds.). Illuminating the Path: Research and Development Agenda for Visual Analytics, IEEE, 2005.
22. Vastenhouw B, Bisseling RH. (2005) A two-dimensional data distribution method for parallel sparse matrix-vector multiplication *SIAM Review*, 47(1), 67-95.
23. VLe: Virtual Laboratory for eScience: www.vl-e.nl
24. Weinan E. , Bjorn Engquist, Xiantao Li, Weiqing Ren and Eric Vanden-Eijnden: *Heterogeneous Multi-scale Methods: A Review*, communications in Computational Physics, Vol. 2, No. 3, pp. 367-450, June 2007.