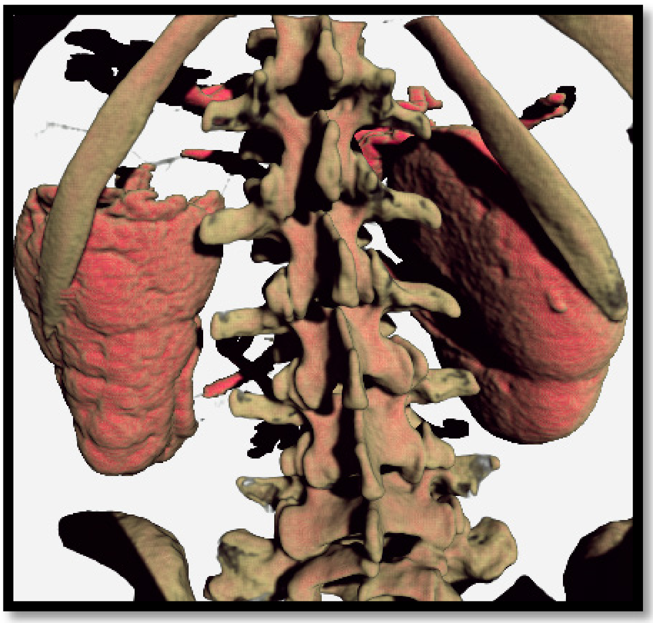
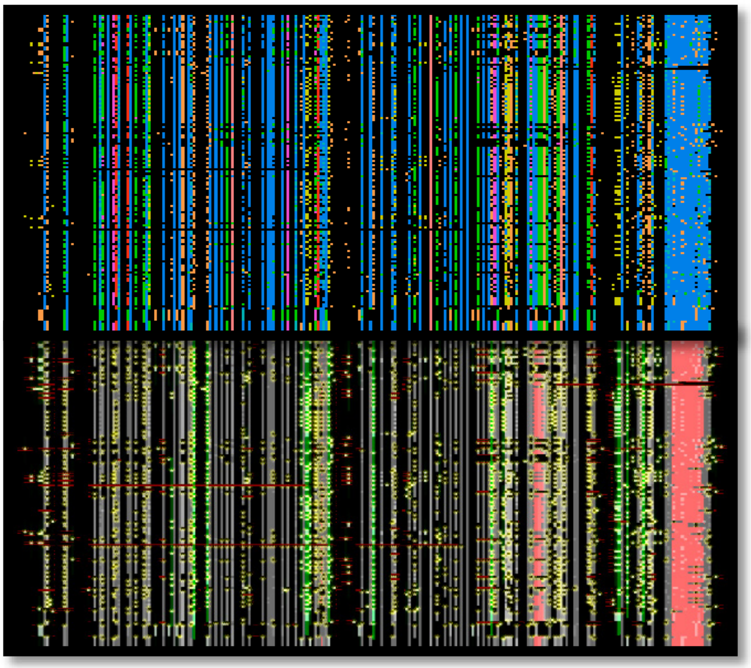
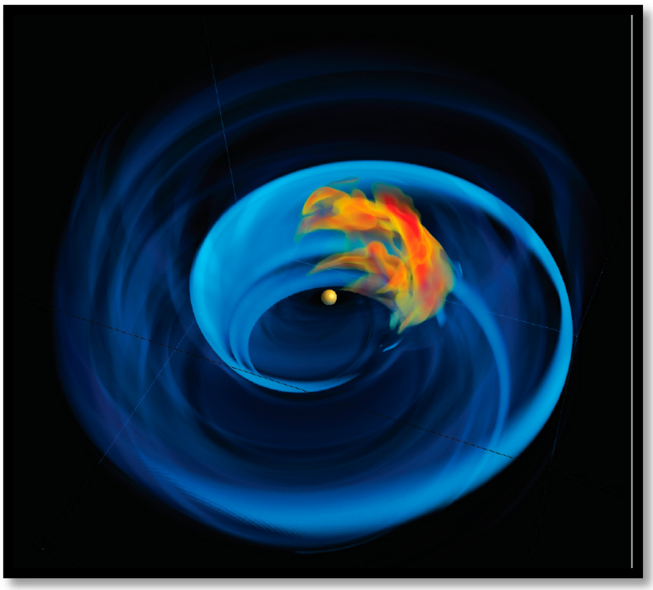
SeRC
Swedish e-Science Research Centre

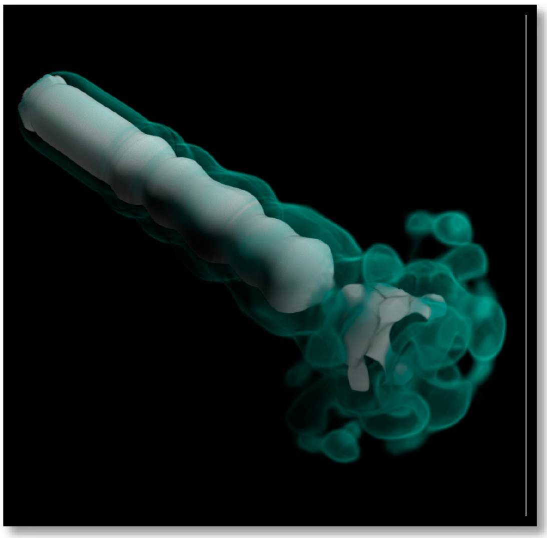
News

2013:1









Visualization Flagship Project

Two new flagship programs, Exascale simulations and Visualization

pages 3–4

SeRC and the Big Data challenge

pages 5–6

Front page: As a core SeRC community, the visualization community deals with data from various sources. The images show a pelvis CT scan, protein multiple sequence alignments, a fuel injection simulation, and space weather fronts (from topleft in clockwise order).

Time to start summing up the first phase of SeRC

This is the third SeRC newsletter and we have now roughly one year left of the first phase of SeRC. During that year the most important activities will be associated with the coming evaluation, both in terms of having great results to report and in terms of making sure that the evaluation itself is prepared in the best way possible. I will count on every one of you to contribute to this process. Should we get a positive evaluation, the funding will become permanent and SeRC will survive for many years to come. We do not know the details of the evaluation yet, but we know that we have to send in a self-evaluation during the fall of 2014, after which there will most likely be a hearing. The decision will be made by the ministry in 2015 based on a report submitted by the funding agencies May 1st of that year. The evaluation will focus on five criteria: scientific excellence, impact, cooperation, education and university efforts.

During the last year the third annual meeting of SeRC took place on April 24 – 25 at the Norrköping visualization center C. About 100 participants attended the conference. Several prominent international speakers were invited, including Hans-Christian Hege from Zuse institute in Berlin discussing data visualization, Björn Engquist from University of Texas giving a seminar about multiscale modeling and Daniel Antoine from British Museum and David Hughes from the Interactive Institute talking about their mummy visualizations. In addition we had a wonderful 3D tour of the universe in the dome theater.

SeSE (Swedish e-Science Education) started during the fall of 2013. It is a collaboration between SeRC and eSSENCE in the area of graduate education, and is built on the two successful graduate schools KCSE at KTH and NGSSC in Uppsala. The SeSE graduate school provides basic training in fields where the use of e-Science is emerging and where education can have an immense impact on the research, but also advanced training for students in fields that are already computer-intensive, see the web page sese.nu for more information.

The increasing recognition of e-Science world wide is well exemplified by this year's Nobel prize in chemistry which was awarded to the "Development of multiscale modelling for complex chemical systems", with the motivation that: "...By taking the chemical experiment to cyber space we can now get answers how chemical reactions work and understand the function of molecules".

An important area that we will focus on during next year is the on-going data explosion which is affecting most scientific and societal areas. It has culminated in the Big Data challenge which demands new tools for storage, analysis and visualization.

In this issue you can read about our activities, like the two new flagship programs, one in exascale simulations and the other one in visualization tools, a new Bioinformatics community project, as well as some exciting industry collaborations. We also discuss the new e-Science report recently presented to the Swedish Research Council and the important area of Big Data.

DAN HENNINGSON
SERC DIRECTOR

New Swedish graduate school in e-Science

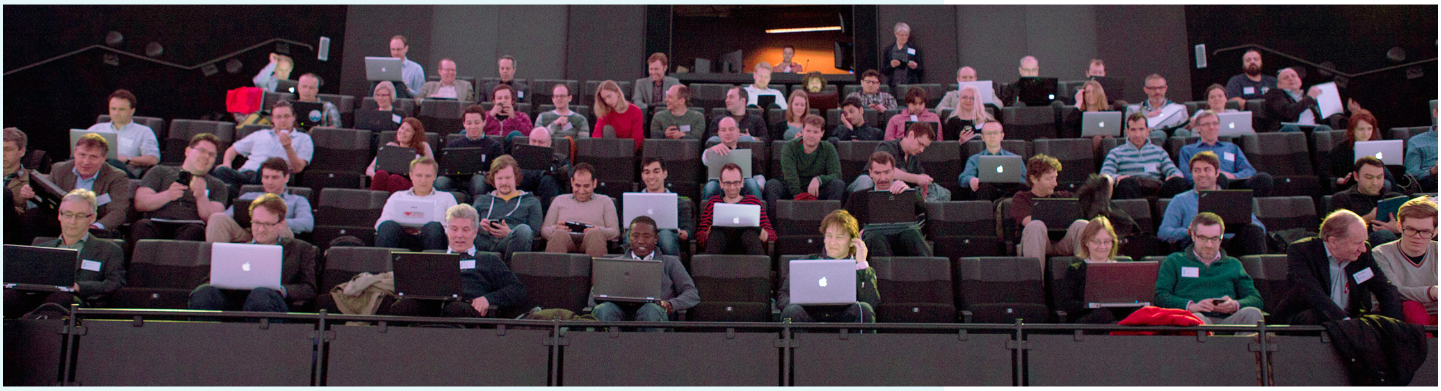
The Swedish Graduate School in e-Science (SeSE) provides courses in the field of e-Science for PhD-students within Swedish academia. For more information see the web page sese.nu.

EASC2014: Solving Software Challenges for Exascale

2–4 April 2014, Stockholm, Sweden

EASC2014 aims to bring together developers and researchers involved in solving the software challenges of the exascale era. EASC2014 is being organized at KTH in association with SeRC and the CRESTA and EPiGRAM projects. More information can be found at http://www.pdc.kth.se/easc2014

The Call for papers for EASC2014 is now open, with a final submission deadline of 12 January 2014. For additional information, please contact Stefano Markidis (markidis@kth.se).



SESSI – The SeRC Exascale Simulation Software Initiati

One of the cornerstones in SeRC is the large number of groups that contribute to the design, development and maintenance of software packages that are used for scientific computing. Over the last few decades we have experienced an amazing development of computer hardware, but it is easy to forget that this would not be particularly useful without huge amounts of hard work to enable simulation software to use it efficiently. With new accelerator architectures the largest machines in the world will likely reach around a billion processor cores around 2024, and already around 2018 we will likely see the first machines that can sustain an Exaflop of computational performance.

SESSI – the SeRC Exascale Simulation Software Initiative – is a new SeRC flagship program with the aim to prepare and support researchers in getting their codes to scale orders of magnitude better than they do today. Several SeRC research groups have already done great work to enable applications such as NEK5000, GROMACS, Dalton and others to use current SNIC resources efficiently. Considering that the investment costs for a single supercomputer can be well over 100 MSEK, and that these applications are used by a large number of users all over the world, improved software performance can directly be translated into enormous financial savings.

Not only is software development more important than ever, it is also more complicated than ever. It is no longer possible for an individual scientist to write a simple code and hope that the compiler will provide good performance – programs will have to be redesigned up with scaling and load balance in mind if they are to run on a Petaflop resource. In addition to this, we are currently faced with a large number of different parallelization techniques such as SIMD instructions, multi-threading, accelerator processors such as GPUs, stream computings or Xeon Phi cards, MPI, but also several new low-level interconnect techniques or special parallel languages.

The program has been inspired by the Petascale Resources Allocation Teams used e.g. at the NCSA supercomputing center in United States, where resources are focused on a small number of widely used applications where the work is expected to have very large impact. In SeRC, a number of researcher teams working with computational fluid dynamics (specifically the NEK5000 code) and molecular dynamics simulations (GROMACS) have initially decided to join forces in this program, and try to learn more from each other when it comes to parallelization techniques. This is also complemented by a new NVIDIA Cuda Research Center awarded to SeRC, and several researchers from the PDC Center for High-Performance Computing that contribute expertise on new parallel programming models, profiling and debugging on complex mixed hardware (e.g. when a code uses both CPUs and GPUs). For each of the codes, this provides us with critical mass in the form of a team of several people that dedicate a large fraction of their time

Collaborative Visual Exploration and Presentation A SeRC Flagship Program in Visual Computing

Visualization plays a crucial role in many SeRC projects. In this flagship initiative, we aim at integrating visualization early on in the discovery process, in order to reduce data movement and computation times.

Within many SeRC reseach projects, visualization is currently often established as the last step of a long pipeline of compute and data intensive processing stages. While the importance of this use of visualization is well-known, facilitating visualization as a final step is not enough when dealing with e-Science applications. As data sets are becoming larger and tasks more compute intensive, the need to integrate visualization early on in the discovery process is of increasing importance, in order to reduce data movement and computation times. Furthermore, this early integration enables an adaption of the visualization algorithms to the data, and intermediate visual results become possible without the need for visualization-based data processing. Within this flagship program we will address the challenges arising from the early collaborations enabled by these in-situ visualizations. We will investigate which role visualization plays to strengthen such collaborations, by enabling a more direct interaction between domain experts within SeRC. The developed concepts are targeted towards the exploration of data, as well as the presentation of the findings. To ensure the relevance of the developed concepts, we collaborate with SeRC researchers from different communities.

Approach

To be able to develop the proposed concepts, it is of uttermost importance to know the demands of the different communities. We have selected a representative subset of communities with which we intensify the collaboration within this flagship program. These communities include:

- Bioinformatics
- Electronic Structure
- FLOW
- Molecular Simulation

The goal is to better understand these domain experts'. Once the concepts developed within this flagship program are satisfying for these communities, the circle of projects will be widened to bring in other SeRC researchers to use and validate the developed methods.

On the long term we expect that the outcome of this flagship initiative has a significant contribution on the scientific discoveries made within SeRC.

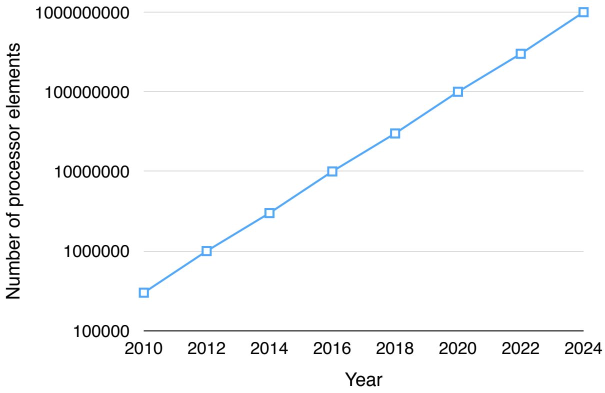
Program PIs

Timo Ropinski

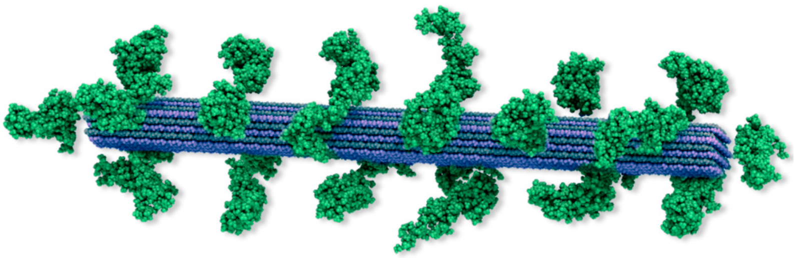
Anders Ynnerman

ve

to enabling it on next-generation hardware. In addition, SESSI also aims to consolidate all efforts related to high-end parallel computation in Sweden, with the goal of establishing a co-design center that other codes will benefit from.



Expected future development of the core-count on the largest supercomputers in the world. The current development is ahead of this projection since the Tianhe-2 supercomputer installed in July 2013 already has 3.1 million processor cores.



Part of a cellulose-lignine system for which molecular dynamics simulations have scaled to more than 150,000 cores with GROMACS. One of the most challenging future tasks is to achieve better scaling for lattice summation algorithms used for long-range electrostatics.

Program PIs

Dan Henningson

Erik Lindahl

Erwin Laure

A community project for SeRC bioinformaticians

The SeRC Bioinformatics community gathers around a common project. The aim is to bring significant progress to this by joining forces from within the diverse community. Within this community project we foresee collaboration with other communities including the Visualization and Parallel and Distributed Computing communities.

Bioinformatics is a diverse field that covers topics from medicine to computer science, and bioinformaticians have backgrounds in math, CS, physics, molecular biology, medicine, and more. SeRC's Bioinformatics community reflects this diversity, and inspired by a SeRC call for collaborative projects, we therefore decided to join forces in a common project.

The starting point of the new project is an ongoing collaboration already funded by SeRC, between Arne Elofsson and Jens Lagergren on protein domain evolution, with a PhD student who works on new mathematical models. This project will be extended in new directions.

The ability to estimate a tree describing evolution given a set of strings representing DNA or protein molecules is taken for granted today and there are many methods and tools available. However, as the understanding of evolution increases, there is an increasing need for models that are more detailed and focus on specific evolutionary features. One such feature is substructures within sequences. Most proteins have a very clear modular architecture of so called domains, essentially recurring sequence patterns or substructures (depending on definition), and there are extensive databases cataloging these domains and their variation.

There is today a good understanding of how domains duplicate, disappear, and combine, to give proteins a wide range of domain architectures. The ambition is now that tailor-made tools, that explicitly models how domains evolve within the proteins, will be produced. This is a natural continuation on earlier work to provide software that infers evolutionary trees describing how genes or proteins have evolved through duplication and loss of genes conditioned on a known species tree. Another layer in the model is now added by detailing how domains evolve within genes/proteins.

This project is a great opportunity to connect to other SeRC communities, like the Visualization community for help on improving the presentation of evolutionary data, and the Parallel and Distributed Computing community, to improve the performance of our computer algorithm implementations.

Two proteins with the leucin-rich repeat being compared by structure. We want to have a better understanding of how these proteins have evolved on the domain level.

SeRC industry collaborations

The interface with industry and society is a cornerstone activity in order to guarantee that the research within SeRC is of strategic relevance. The work to find suitable forms of collaboration that are of mutual interest will be intensified during 2014.

One of the success stories so far in the outreach to industry is the collaboration with Exotic Matter, a small company that developed software used to create visual effects of fluids for the movie industry. Generating realistic fluid effects is hard and time consuming since it requires solving the actual equations describing the fluid motion mathematically, namely the Navier-Stokes equations. A master thesis project between Exotic Matter and the SeRC FLOW community was defined in order to investigate the potential of using GPU's to speed up the software and thus lower the turn-around time to generate a movie sequence.

Scania is a good example where the collaborative efforts have led to substantial knowledge transfer both in terms of high performance computing, large data storage but also an increased understanding about the challenges and demands industry are facing and how to meet these demands. To aid the knowledge transfer common workshops have been arranged and a number of possible master thesis projects have been identified. So far two master thesis projects have been executed.

Sectra is one of today's major diagnostic imaging providers, and has its roots in research performed in collaboration with the Visualization community at Linköping University. Sectra continue to have intense collaborations with SeRC reserchers. Throughout the SeRC communities there are numerous other collaborations on different levels with external partners such as industries, research institutes, public agencies, SME's and more. Some of these collaborations are new and some of them existed before SeRC was formed.

Some of our industrial collaborators are BinaryBio, Exotic Matter, FOI, Light-Lab of Sweden, Nanologica, Portendo AB, SAAB Group, Scania and SMHI.

Visual effects of fluids for the movie industry produced by Exotic Matter

SeRC and the Big Data challenge

Many scientific and societal fields, from high-energy physics to Facebook, have seen an explosion in data in recent years, resulting in data sets so large and complex that it is difficult or even impossible to store or process them with traditional methods. The term *Big Data* refers to datasets that require new tools for storage, analysis and visualization. The definition first appeared when data sets at Google and Amazon grew so large that they no longer could be handled in traditional databases. This led to the development of new generations of tools to interact with this data and mine it for new information and correlations. SeRC researchers are already very well established in many fields where Big Data issues have to be dealt with and it is an important goal to further expand Big Data collaborations for the strategic research area. This will be addressed by several activities during the next year, among them an emphasis on Big Data aspects at the next annual meeting. We briefly mention below some of the Big Data related activities SeRC researchers are involved in.

- SeRC bioinformatics researchers at the Science for Life Laboratory have been instrumental in the analysis that lead to the sequencing and publication of the world's largest genome to date – the Norwegian spruce. SeRC also supports work on automated recognition of peptides in mass spectroscopy by using new machine-learning techniques for matching to gigantic databases.
- SeRC bioinformatics researchers at Linköping develop automated methods and tools to annotate and describe data in terms of ontologies for use in management and integration of data.
- The flagship project eCPC (e-Science for Cancer Prevention and Control) exemplifies the Big Data concept in biomedical research. A data-availability framework is developed for jointly addressing Swedish biobanks and molecular sequence data, population register data and data from national quality registries (clinical patient-related data). In addition secure protocols for sensitive data in distributed environments, including safe utilization of cloud computing resources is studied. The purpose of eCPC is to use the complex data sources to set up prediction tools and through modeling and simulation extract *in silico* how different screening strategies affect population mortality, morbidity, side effects and cost.
- A further example of Big Data related research are biobanks. Jan-Eric Litton (SeRC steering group member) has been appointed Director General of BBMRI-ERIC, a European collaboration between biobanks that is now implemented within a European Research Infrastructure Consortium (ERIC). Big Data aspects of biobanks are also dealt with in the SeRC-related FP7 project BiobankCloud that focuses particularly on secure and efficient biobank data storage and analysis facilities.
- The molecular simulation community has led the development of new techniques to use millions of loosely coupled distributed computing simulations and use new analy-

sis methods to automatically detect weak correlations. Through a SeRC-initiated collaboration with PDC this has resulted in several EU-projects dealing with the exchange and annotation of large-scale simulation data.

- The Visualization community focuses on the challenges related to the visualization of large datasets and develop algorithms for interactive visualization of such data. A new Flagship program with the goal to further strengthen the interdisciplinary collaboration between the visualization community and the applied communities in relation to visualization of such large datasets has just been started.
- Several projects in SeRC's Distributed and Parallel Techniques community deal with efficient data storage and analysis frameworks, building on and extending tools like Hadoop, Spark, and Pregel. Frameworks are being field tested with pilot applications from bioinformatics, biobanks, the social sciences, and industry (e.g. Spotify).
- Data is also produced from very large simulations in climate and turbulence research, and here too the fields are faces with entirely new challenges for data management, integrity, analysis and visualization. The SeRC initiative has made it possible to support new such collaborations.

Cases for Swedish e-Infrastructure presented to the Swedish Research Council

During 2013 an investigation with the overall mission to provide the Swedish Research Council with executive information on the scientific requirements on future e-Science infrastructures in Sweden has been conducted. The resulting report was presented to Council for Research Infrastructures at the beginning of November. The report describes selected cases that provide examples of the scientific results that can be obtained if the specified infrastructure requirements are met. The work on documenting the e-Science cases was conducted by seven panels consisting of leading experts within selected research areas.

The cases are described from a scientific perspective and potential breakthroughs that can be enabled by the use of a future e-Infrastructure are described for each case. The demands put on the e-Infrastructure are then extracted. The reports presents evidence showing that a significantly increased level of investment in e-Infrastructure is required to enable research leading to the described breakthroughs. The report emphasizes the need for investments not only in hardware infrastructure but also shows that the software and human infrastructure must be given priority. The most urgent needs for the community are described in terms of:

- Capacity computing – Dedicated special purpose HPC systems.
- Storage – Large scale storage solutions that are integrated with database and visualization services.
- Software – Efforts to develop new software to address new problems and new approaches.
- User support – The pool of human resources providing qualified assistance to users.

The report also underlines the need for increased complimentary efforts on data driven research and supporting services, such as policies and legal frameworks.

ANDERS YNNERMAN
SERC CO-DIRECTOR

SeRC Faculty and Steering-group strategy-meeting in October together with Perlan consultants.

Fifth Annual Meeting of SeRC

The fifth annual meeting of SeRC will take place 23th – 24th of April 2014, in Stockholm.

ROYAL INSTITUTE OF TECHNOLOGY

Stockholm University

Linköping University

Karolinska Institutet

Printed by LiU-Tryck, Linköping

Swedish e-Science Research Centre,
SE-100 44, Stockholm, +46 8 790 7570, info@e-science.se
Editor: Olivia Eriksson For a paper version or PDF-file send an e-mail to info@e-science.se.