Technology-enhanced research

Dr Sandra Gesing develops computational tools that enable scientific researchers to analyse big data via high-performance computing more simply, efficiently and sustainably

Could you start by explaining how you developed an interest in computer science, and what motivates you in your research?

Computer science was offered as a class at my high school. Besides software and hardware theory, we were taught programming, which became one of my hobbies. I decided to make my hobby my career and completed an apprenticeship in computer science. While I enjoyed working in industry as an administrator, system developer and head of a systems programming group over the following 12 years, I realised early on that I would like to delve deeper into computer science and I did my German diploma (equivalent to a Master's degree) via extramural studies.

I then became especially interested in the important and challenging topics that bioinformatics supports, such as drug design and eradicating diseases, and I started to work in academia in a bioinformatics group. My PhD was focused on science gateways for molecular simulations. From this, I developed a primary interest in science gateways, workflows and distributed computing. While I know that I won't find cures for diseases or eradicate them, my skills enable me to contribute to projects with these goals.

Can you explain why science gateways are so important for advancing scientific discoveries?

Science gateways are software solutions – now often accessible via a web browser – tailored to the requirements of a specific community while hiding complex underlying infrastructures. They are important for enhancing science, since researchers are experts in their disciplines but not necessarily information technology specialists.

What are some of the biggest achievements of your research career to date?

Because of the interdisciplinary nature of my research, a large network of collaborators is essential for gathering ideas and understanding the pain points of researchers using complex computing infrastructures. The three biggest achievements in my research career are the successful MoSGrid science gateway (an intuitive portal for the molecular simulation community), a large international network of interdisciplinary collaborators and the European International Workshop on Science Gateways (IWSG) I founded in 2009 and have guided since.

Successful collaboration in MoSGrid is reflected in over 20 scientific publications. I led the design and implementation as work package leader of the project. IWSG is the partner workshop series of the US science gateway workshop series, as well as the Australasian science gateway series. Each year, IWSG attracts 30-60 participants, and it has an active community and enjoys strong support from collaborators.

You are particularly interested in science gateways for bioinformatics applications. Why is this an exciting area?

Bioinformatics is an exciting area for me for two reasons. First is the wide range of important applications in health and life sciences. The second is its novel technologies, such as next-generation sequencing, which allow data creation in exascale dimensions with relatively minor effort and monetary resources compared to only a decade ago. We are able to find answers for research that would not have been feasible before – maybe the questions were not feasible to ask before. But the amount of data creates new challenges, which obviously need new software solutions.

Have you faced any major challenges?

The major challenge is the breadth of topics associated with science gateways and workflows. While my field is exciting for this reason, the challenges are manifold, from intuitive user interfaces and security features to efficient data and workflow management, and parallelisation of applications employing parallel and distributed computing architectures. I overcome these challenges by becoming acquainted with the target domain and applying novel solutions. Close collaboration with researchers from the target domain and calling on experts for specific aspects, such as librarianship, statistics and machine learning, are essential, since it is not practicable to become an expert in every aspect of a science gateway.

Why are you so passionate about enhancing the reproducibility of science?

Reproducibility is a cornerstone in science for validating results. Research applying computational simulations and methods is predestined to support reproducibility in an easy way. Science gateways and workflow systems are promising vehicles to achieve this goal.

Could you outline some solutions that you have been working on recently?

One current solution is a prototype for easing the use of a software for calculating the dynamics of mosquito populations by exploiting graphical processing units. This software could be used for malaria research, for example. Another solution we are working on is concerned with combining two workflow systems to make use of the strengths of each and increase computational efficiency.

Gateways to advancing science

The Center for Research Computing at the **University of Notre Dame**, USA, is developing computational tools that enable the creation, distribution and widespread use of vital scientific knowledge

AS SCIENTIFIC KNOWLEDGE has developed, so have the complexities of the problems researchers seek to solve. Discovery and innovation in science increasingly depend on multidisciplinary knowledge. Computation has therefore become a fundamental tool, not only for developing and testing hypotheses but also for streamlining collaboration across disciplinary, organisational and geographic boundaries.

Leveraging the massive data processing capabilities offered by parallel, grid and distributed computing, as well as internet services, is essential for harnessing the power of big data generated by scientific research projects. Dr Sandra Gesing, who is based at the University of Notre Dame in Indiana and is the founder of the European Science Gateways International Workshop series, is confronting these challenges head on. Her research predominantly serves the bioinformatics domain, and recent projects have addressed a variety of technical and logistical issues, such as improving both computational performance and user access to accelerated architectures.

"Gesing draws on her experience in developing quality software and her deep understanding of computational strategies to open up new opportunities; working with experts in other disciplines, she advances computational technology until it meets the challenges posed," states Professor Malcolm Atkinson of the University of Edinburgh, who is one of Gesing's collaborators.

SCIENCE GATEWAYS

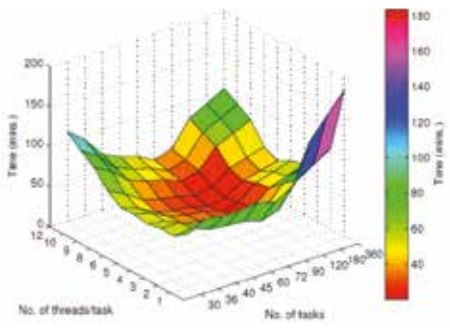
For scientists without extensive knowledge of information technology, it can be difficult to use a portal involving terabytes or petabytes of data and teraflop or petaflop computing power to obtain results, especially when time is of the essence. A science gateway is a system designed to make the task easier, based on a distributed computing infrastructure and delivering integrated access to specific sets of data and accompanying analysis and modelling tools, via a web portal or desktop computer application.

"Gesing's science gateway research aims at designing efficient methods for connecting scientific applications to cyber infrastructure, in the end making scientific discoveries faster and easier," points out Jarek Nabrzyski, Director of the Center for Research Computing. Key to this is delivering a seamless, intuitive user experience, making it feasible for any scientist to extract information, run workflows, execute simulations and develop complex algorithms without the need for a deep understanding of the underlying computing infrastructure.

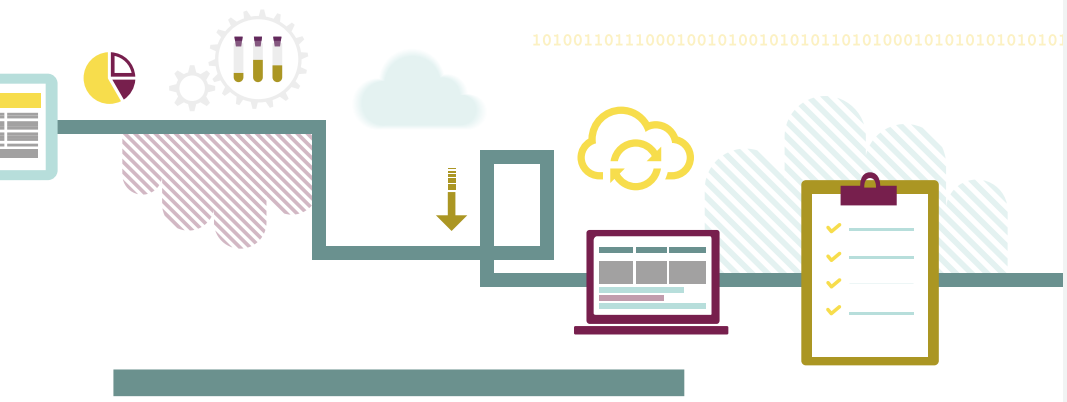
EASY-TO-USE COMPLEX RESOURCES

Workflow software products are useful to scientists for analysing observational and experimental data, and for running models and simulations. However, their effective use requires knowledge of their underlying technologies, so the learning curve is steep. In addition, support for the whole workflow lifecycle and adaptability tends to be poor. Thus, in recent projects, Gesing has developed frameworks to help scientists manage workflows from end to end while making optimal use of massive data mining and computing resources.

Where possible, Gesing's frameworks reuse existing software and procedures. She then enhances them by applying agile web methods, state-of-the-art infrastructure design and deep knowledge of the abilities and attributes of existing science gateway technologies that exploit research infrastructures such as the US Extreme Science and Engineering Discovery Environment (XSEDE) and the Partnership for Advanced Computing in Europe (PRACE). A central feature is the use of application programming interfaces (APIs) to enable rapid tailoring of the system to the end user's data, task and/or workflow requirements. APIs enable information technology developers to create a solution without needing to start from scratch. For users, APIs enable access to workflow systems through a single graphical user interface, allowing them to easily edit and



Impact of multilevel concurrency (thread-level parallelism) and distributed computing (task-level parallelism) on the execution time of a data-intensive workload. Selecting a good runtime configuration can optimise resource utilisation. (O. Choudhury, et al. 2015. *Cluster Computing (CLUSTER), 2015 IEEE International Conference on.* p390-393.)



Gesing anticipates that improving the usability and sustainability of science gateways will enhance the reusability and reproducibility of scientific research

monitor workflows and simultaneously optimise their usage of data from one screen.

Gesing developed her graphical user interface design through close collaboration with scientists. It is based on a web browser 'dashboard' concept, which connects seamlessly to both the workflow system and external resources like Cloud infrastructures, meaning the user does not need to become acquainted with diverse interface layouts. The dashboard can be tailored to each user, allowing researchers to focus on their own research topics instead of trying to familiarise themselves with a complex data and computing infrastructure.

IMPROVING ACCESS TO RESEARCH DATA AND PUBLICATIONS

Gesing anticipates that improving the usability and sustainability of science gateways will enhance the reusability and reproducibility of scientific research. Unfortunately, despite living in a digitally connected world, access to research data and publications remains difficult, and is often restricted by local copyright and funding policies. If the results of prior research projects were made more readily available, this would allow other researchers to build on them rather than simply reinventing the wheel.

Using the German Molecular Simulation Grid science gateway as a case study, Gesing recently examined the possibilities of integrating different research infrastructures. This has led her to advocate for the examination of policies governing each research infrastructure. Only when policies are aligned can technical integration follow.

DRIVING OUTREACH, ENGAGEMENT AND EXPERTISE

Gesing places strong emphasis on collaboration and she works closely with a number of international researchers. Dr Michelle Barker, Deputy Director of National eResearch Collaboration Tools and Resources and founder and co-chair of the International Workshops on Science Gateways – Australia, speaks warmly of their cooperative efforts: "My collaboration with Gesing has expanded my vision of what is possible in my research programme, as her

The concept of the Science Gateway Institute.

suggestions on alternative approaches and practices have helped elevate the programme to one that provides leadership in this field," she says.

Gesing is also heavily involved in a proposed US Science Gateway Community Institute, where her role focuses on outreach and community engagement. "Our proposal for the National Science Foundation (NSF)'s software institute programme will help advance both science gateways research and outreach to diverse research communities. We envision the Institute will help the research community by providing a range of services to help them develop science gateways," said Nancy Wilkins-Diehr, Associate Director at the San Diego Supercomputer Center, lead Principal Investigator of the SGCI – and one of Gesing's key collaborators. "One important part of this is carving out career paths for gateway developers and economies of scale for academic institutions achievable by creating gateway expertise groups on campuses."

For the foreseeable future, Gesing has her work cut out for her. She is now producing a guide to best practice in science gateway development and is lead editor for a forthcoming book about bioinformatics and big data. She is also participating in an ambitious worldwide project – the International Consortium for Technology in Biomedicine – with partners in the US, Europe and India, which is designing science gateways for biomedical applications. Moreover, together with Barker and Wilkins-Diehr, she is also planning to establish a global coalition on science gateways.

SCIENCE GATEWAYS, WORKFLOWS AND DISTRIBUTED COMPUTING

OBJECTIVE

To develop computational strategies that efficiently connect scientific applications to cyber infrastructure in order to accelerate scientific discoveries.

KEY COLLABORATORS

Jarek Nabrzyski; **Scott Emrich**; **Douglas Thain**; **Natalie Meyers**; **Ian Taylor**, University of Notre Dame, USA • **Thomas Connor**, Cardiff University, UK • **Malcolm Atkinson**, University of Edinburgh, UK • **Rizos Sakellariou**, University of Manchester, UK • **Michelle Barker**, Nectar, Australia • **Nancy Wilkins-Diehr**, University of California, San Diego, USA • **Maytal Dahan**, The University of Texas at Austin, USA • **Katherine Lawrence**, University of Michigan, USA • **Marlon Pierce**; **Suresh Marru**; **Sudhakar Pamidighantam**, Indiana University, USA • **Michael Zentner**, Purdue University, USA • **Linda B Hayden**, Elizabeth City State University, USA • **Jens Krüger**, University of Tübingen, Germany • **Peter Kacsuk**, MTA SZTAKI, Hungary • **Tamas Kiss**; **Gabor Terstyanszky**, University of Westminster, UK • **Richard Grunzke**, Technische Universität Dresden, Germany • **Sonja Herres-Pawlis**, RWTH Aachen University, Germany • **Johan Montagnat**, CNRS, France • **Horacio Perez Sanchez**, University of Murcia, Spain • **Jesus Carretero**, Universidad Carlos III de Madrid, Spain • **Gabriele Pierantoni**, Trinity College Dublin, Ireland • **Ivan Merreli**, CNR-ITB, Italy • **Cezary Mazurek**, Poznan Supercomputing and Networking Center (PSNC), Poland

PARTNERS

IEEE Technical Committee on Scalable Computing

FUNDING

National Science Foundation (NSF) • National Institutes of Health (NIH) – no current funding but past fundng and a grant under review • Bill & Melinda Gates Foundation – past funding

CONTACT

Dr Sandra Gesing
Research Assistant Professor

Center for Research Computing
PO Box 539, Notre Dame, Indiana 46556, USA

T +1 574 631 2400
E sandra.gesing@nd.edu

https://engineering.nd.edu/profiles/sgesing

http://sandra-gesing.com

www.researchgate.net/profile/Sandra_Gesing

www.linkedin.com/in/sandragesing

@sandragesing

SANDRA GESING is a research assistant professor and computational scientist at the University of Notre Dame, USA. Previously, she worked as a research associate at the University of Edinburgh, UK, and at the University of Tübingen, Germany, where she also received her PhD. Additionally, she has extensive experience working as a project manager and developer in industry.

CRC
CENTER FOR RESEARCH COMPUTING