

COLLABORATIVE Science in the Age of



Tony Hey, formerly director of the UK eScience initiative, is now the Corporate Vice President of External Research at Microsoft.

time working on technology. Of course, while the potential exists, there is still a lot of work to be done.

What's in a name? You held the post of Director of the UK eScience Core Program for four years. Would you explain the distinction between eScience and the Grid? Or are the two just different names for distributed computing?

To me, eScience is about the use of software technologies for solving scientific problems. It is about developing a set of tools that will support the data-centric, networked, collaborative, and often multidisciplinary science that is needed to address the next generation of scientific challenges. It is about giving researchers tools that enable them to spend more time in doing research instead of struggling with computer and infrastructure issues. It is about enabling them to handle the deluge of scientific data and, in the process, allow them to address grand challenges, to ask and then answer more interesting questions.

The Grid is just one of the tools in the arsenal of a researcher. It is a suite of distributed computing technologies that allows the easy sharing of computing resources between organizations. The Large Hadron Collider (LHC) particle physics accelerator at CERN (the European Organization for Nuclear Research) will generate petabytes of data, and that data and its analysis must be distributed across the large, global, experimental collaborations. When the LHC starts generating significant amounts of data next year, the particle physics community at the LHC will be a "production" demonstration of how Grid computing technologies can be deployed across organizations to process the huge amounts of data and computation. The TeraGrid effort in the U.S. is another example where large-scale computational resources from around the country are brought virtually together and offered to researchers from easy-to-use portals.

Cloud Computing is another important paradigm that is now emerging. This has a different model for making resources available to eScientists around the world. The domain is still in its infancy, but big companies are leading the way in becoming providers of utility computing (for example: Microsoft, Google, Amazon). At Microsoft, we see great potential in the use of Cloud Computing in eScience, and we are making sure that our commercial and research investments reflect that vision.

SciDAC Review: *You have worked most of your life in academia and government. What is it like to make the transition to working at Microsoft?*

Tony Hey: After helping the UK eScience Program get off the ground, I thought really hard about my next challenge. I truly believed in the vision and goals of the eScience program and really enjoyed seeing the investments in people and projects paying off.

When I started the UK eScience Program in 2001, it never occurred to me that Microsoft could play a key role in eScience. However, after many discussions with Jim Gray, I realized that I was wrong. When Craig Mundie, chief research and strategy officer for Microsoft, shared with me his vision for a "Technical Computing" research activity at Microsoft, I realized the great potential for making a huge difference in the way research is done worldwide. Microsoft Research represented a great opportunity to take the eScience vision and realize it at a global scale.

I believe that Microsoft is perhaps the only company with the technology depth to touch every aspect of the research lifecycle—from data acquisition, management, analysis, and visualization to scholarly publishing, knowledge dissemination, and archiving. Thus Microsoft is in a unique position to bring together its ecosystem of tools and services and make them relevant to researchers all around the world. This will help enable researchers to realize their full potential by allowing them to concentrate on what they know best, their research discipline, rather than spending

ULTRASCALE Computing

In a meeting at the Institute of Materials in 2002 you mentioned that “many areas of science and industry rely on the collaboration between centers of expertise, data, and computing around the world.” Can you give one or two examples of successful collaborations?

As former Microsoft Researcher Jim Gray once observed, science today is becoming predominately data-centric (instead of experimental, theoretical, or computational). Scientists have to deal with such a deluge of data and information that makes the use of powerful technologies to automate the management, mining, and analysis of data increasingly important. The large data management and processing requirements impose

To me, eScience is about the use of software technologies for solving scientific problems. It is about developing a set of tools that will support the data-centric, networked, collaborative, and often multidisciplinary science that is needed to address the next generation of scientific challenges.

significant demands on the computer-based infrastructure and, in the future, may well increase the requirements to levels where no single institution can afford to build and support it. The information technology (IT) industry can help address that concern by significantly bringing down the cost of the infrastructure because of their economies of scale. For example, Amazon can offer compute and storage resources on demand at a fraction of the cost of an institution-based solution. That's the promise of Cloud Computing.

On the software side, industry can partner with the academic open source community to build product-quality, robust, well-documented software that can be reliably deployed to support large infrastructures.

Some of today's “Grand Challenge” research problems and the eScience infrastructure and tools that we are helping to develop will be relevant to industrial problems of tomorrow. A prime example is bioinformatics, genomics, and proteomics. The technologies being developed by researchers to understand fundamental scientific problems of biology and disease will be increasingly relevant to the pharmaceutical and healthcare industries. In the process of drug design and testing, biologists have to work with chemists and genomicists and doctors who are in different departments or may be located on different sites. The tools and technologies from eScience will become fundamental to the whole of the pharmaceutical industry and help reduce the cost of new drug development for the benefit of the whole world.

A key feature of the UK eScience Program is its engagement with industry. What were some of the major accomplishments in this area during your tenure as director?

The eScience Program had significant involvement from industry from the start since part of the funding for the Core Program that I managed came from the Department of Trade and Industry. Besides the major IT companies there was participation from nearly 100 other companies, both large and small. For example, the Distributed Aircraft Maintenance Project (DAME) demonstrated how university researchers could work together with engineers from Rolls-Royce to build monitoring solutions for aircraft engines as part of the ever continuing efforts to improve reliability and reduce costs. Besides Rolls-Royce the project included two other companies—Data Systems and Solutions and a spin-off company from the University of York, Cybula, specializing in neural network technology.

Another example is the DiscoveryNet project led by Yike Guo from Imperial College, London, which investigated workflow technologies for bioinformatics and pharma. This project led to a successful Imperial College spin-off company called InforSense.

A related question: what role do you see collaborative computational science playing in the major transformations on the horizon for energy sources and technologies?

Some of our research projects are looking at major environmental issues such as our work with the Berkeley Water Center and the FluxNet and Ameriflux communities.

We also have a project in Switzerland—the Swiss Experiment—that is currently instrumenting the glaciers of the Swiss Alps. Since the snow-pack in the Alps supplies much of Switzerland’s water, this is an important research area for the country and the region.

I also think that we are entering a new era where a model combining “software + services” will be the norm for both business and science. Cloud Services will rely on vast data centers containing many thousands of servers and consuming large amounts of energy for both power and cooling. Like other IT companies, Microsoft is looking at the energy efficiency of its data centers, and my External Research organization has just completed making awards from our RFP in “Sustainable Computing”. In addition, it is exciting to me personally that Dan Reed has recently joined Microsoft Research. Besides working with us on the multicore research agenda, he will be leading our research into next-generation data centers.

Now that petascale is a real possibility, exascale computing is next. What do you believe are the biggest opportunities for collaborative science in the age of ultrascale computing? The biggest obstacles?

The need for collaboration between scientists will not change, no matter how large the available computational resources. The need for collaboration will always exist as long as the scientists keep asking bigger and bigger questions

The need for collaboration will always exist as long as the scientists keep asking bigger and bigger questions and as long as the grand challenges continue to get addressed and replaced by even greater ones.

and as long as the grand challenges continue to get addressed and replaced by even greater ones.

At Microsoft we are actively engaged in the emerging petascale revolution. While petascale supercomputing is just around the corner, it is still the case that only a few will have access to it. In the computing arena Microsoft is looking at the opportunities that exist for democratizing the huge computational power that the multicore computing era will bring. We believe that the next few years will further transform the way we do research. Researchers will have access to computers that are capable of performing orders of magnitude faster than the current generation and with compute resources in the cloud that cannot be imagined today. Microsoft and Intel recently announced joint funding of \$20 million for two centers at the University of California–Berkeley, and the University of Illinois–Urbana-Champaign to do research in the multicore computing

domain. The goal is to significantly advance the state of the art in system architecture, runtime environments, programming languages, and tools, and to generate mass-market multicore applications. In addition, Microsoft has also funded multicore research projects at Indiana University, Rice University, and the University of Tennessee, as well as a Multicore Research Center with the Barcelona Supercomputing Center.

I also believe that the availability of large computational power will be accompanied by a further explosion of data. The big challenges will still be associated with the movement of that data (network bandwidth) and its management (tools to process, analyze, visualize, and reason over it). We have a number of projects addressing the challenges related to data, information, and knowledge.

Systems biology has been called the science of the 21st century. Has eScience gotten involved in this multidisciplinary and highly-collaborative area?

Yes, in several ways; we are working on several biology projects that address core eScience challenges. One of these is the Connectome Project at Harvard University which has the long-term goal of constructing a wiring diagram of the human brain—with an estimated 300 million synapses per cubic millimeter of brain, this is an immense challenge. Work is in its early stages, but we have been able to help researchers navigate the immense volume of visual data being generated. Automated data collection requires standardization and processing through workflows. In a different project with the team of Professor Carole Goble at Manchester University in the UK, we are looking at the integration of the myExperiment project with Microsoft desktop tools.

The challenge of biology is its complexity and computing has a lot to learn—by attempting to model living systems, for example. We have a team in Microsoft Research Cambridge working in this area to design new programming languages specifically to model biological systems, and applying methods more familiar to computer scientists than biologists to analyze these systems to gain new biological insight. These researchers work closely with bench biologists to validate their results, and we hope this type of work will benefit both biology and computing in the future. We have good evidence that this is happening already in a different project. Our work with Riccardo Zecchina at the Politecnico di Torino has resulted in the creation of new algorithms that can be applied equally well to the management of computer networks as to the inference of reaction networks inside a living cell. Whether it is large-scale data storage or the challenges of data integration, modeling, analysis, and collaboration, biology has a lot to teach us about computing, and vice versa. But of course these challenges are not unique to biologists, and we are also looking at more general portals and tools that can be adapted to different scientific domains.

You co-authored an article in Science in 2005 titled “Cyberinfrastructure for eScience,” in which you envisioned vast distributed digital repositories of scientific data that would enable

Biography in Brief: Tony Hey

As Corporate Vice President of the External Research Division of Microsoft Research, Tony Hey is responsible for the worldwide external research and technical computing strategy across Microsoft Corp. He leads the company's efforts to build long-term public-private partnerships with global scientific and engineering communities, spanning broad-reach and in-depth engagements with academic and research institutions, related government agencies and industry partners. His responsibilities also include working with internal Microsoft groups to build future technologies and products that will transform computing for

scientific and engineering research. Hey also oversees Microsoft Research's efforts to enhance the quality of higher education around the world.

Before joining Microsoft, Hey served as director of the UK eScience Initiative, managing the government's efforts to provide scientists and researchers with access to key computing technologies. Before leading this initiative, Hey worked as head of the School of Electronics and Computer Science at the University of Southampton, where he helped build the department into one of the pre-eminent computer science research institutions in England.

Hey is a fellow of the UK Royal Academy of Engineering and a member of the European Union's Information Society Technology Advisory Group. He also has served on several national committees in the United Kingdom, including committees of the UK Department of Trade and Industry and the Office of Science and Technology.

For his service to science, Hey received the award of Commander of the Order of the British Empire in the 2005 UK New Year's Honours List.

Hey is a graduate of Oxford University, with both an undergraduate degree in physics and a doctorate in theoretical physics.

a "new generation of collaborative science software applications." How close are we to realizing this vision?

We are already seeing lots of scientific disciplines ready to share their data. In astronomy, for example, we have the Sloan Digital Sky Survey/SkyServer data repository, which was funded by a large number of collaborating organizations. A new survey project led by the University of Hawaii's Institute for Astronomy is the Panoramic Survey Telescope And Rapid Response System (Pan-STARRS). This will generate a map of the sky of unprecedented quality, and the petabytes of data it generates will be available for all scientists and enthusiasts to access. We are also seeing visualization tools such as Microsoft Research's World-Wide Telescope (free, public beta version made available in May 2008) assist users in building wonderful experiences on top of such shared digital repositories.

Repositories in other fields of science are also becoming increasingly important. The National Library of Medicine, for example, stores and curates many different databases from a literature repository like PubMed Central to a data repository like PubChem. These repositories and others like them will continue to grow and more will emerge. Even Microsoft has entered this space with our recent preview of our Research Output Repository platform. The digital platform builds on robust product technologies such as SQL Server and the .NET Framework and will be offered free to communities wishing to build their own digital repository solution.

One of your responsibilities as Corporate Vice President of External Research at Microsoft is building long-term partnerships with global scientific and engineering communities. What particular challenges do you face in carrying out this activity?

It is my belief that computer science can be a key driver for advancing the many fields where computer technology

intersects with traditional sciences such as in astronomy, biology, medicine, and engineering. The computer science research technologies vary significantly—from machine learning algorithms for fighting AIDS and database technologies for hydrology to new pedagogies for teaching/education and XML-based formats for the archiving and preservation of information.

Computer science can be a key driver for advancing the many fields where computer technology intersects with traditional sciences such as in astronomy, biology, medicine, and engineering.

My job is exciting because I have the opportunity to work with multiple product and research teams in Microsoft as well as with leading researchers in many different areas of science. I believe that we are uniquely positioned to help bring together the resources the tools, infrastructure, expertise, and support for collaborative environments that can accelerate game-changing research across the globe. My goal as Director of External Research in Microsoft Research is to make sure that Microsoft offers an open and creative set of technologies that allow researchers to have the flexibility they need to be innovative in their area of science.

Thank you for taking the time to answer our questions.