CRD

Computational Research Division Report

Berkeley Lab

Office of Science
U.S. Department of Energy

DECEMBER 2007

Deconstructing Microbes

Metagenomic research on bugs in termites relies on new data analysis tools

A metagenome data management and analysis system developed and managed by CRD researchers contributed to a breakthrough research recently published in Nature that probed the genetic materials of microbes in a termite's gut.

Understanding how these microbes convert wood to energy for sustaining a termite's life would help researchers figure out better ways to convert biomass to biofuels in factories. The microbes exude enzymes that efficiently break down the wood's cellular structure, but scientists so far have little understanding of how the process works.

The research, led by the DOE Joint Genome Institute (JGI), relied on IMG/M, a system developed for managing and analyzing metagnome data. Metagenomics is a study of genetic material collected directly from organisms in their natural environment, a relatively new and alternative approach to extracting genetic information *continued on page 4*

Scientists study the ability of microbes residing in these termintes' hindguts to break down the wood's cellular walls and convert them into sugars. The research relies on a megatenomic data management and analysis system called IMG/M, developed by the Biological Data Management and Technology Center in CRD.

Popular Science

A nanomaterial research paper in Nano Letters drew strong interest from the scientific community

A research paper showing how zinc oxide can be manipulated to become a good material for photovoltaic devices was among the most-accessed paper published by Nano Letters in the third quarter.

The paper, written by Joshua Schrier, Denis Demchenko and Lin-Wang Wang in CRD's Scientific Computing Group, laid out the calculations that narrowed the band gap — the energy difference between the top of the valence band and the bottom of the conduction band — of zinc oxide. To effectively absorb light, materials must have band gaps which match the distribution of photons that pass through the earth's atmosphere, from the sun.

Nano Letters, published by the American Chemical Society, recently unveiled the 20 most downloaded or viewed papers from July through September. Schrier's paper, "Optical Properties of *continued on page 3*

DIRECT APPROACH

Researchers employ an algorithm to solve an energy-reduction issue essential in describing complex physical system

The SIAM Journal on Scientific Computing recently published the work by three CRD researchers that incorporated an algorithm they developed in computing the ground state energy and wave functions for describing the mechanics of a complex physical system.

The paper, "A Trust Region Direct Constrained Minimization Algorithm for the Kohn-Sham Equation," described the use of the direct constrained minimization (DCM) algorithm with the "trust region" technique to compute single electron wavefunctions associated with the ground state energy of molecules or solids. These wavefunctions are also solutions to the Kohn-Sham equations, a set of equations that form the first order necessary condition satisfied by the minimizer of the total energy functional.

The Kohn-Sham equations are traditionally solved by a technique called the self-consistent field (SCF) iteration. However, the simplest form of SCF often fails to converge. The convergence of SCF can be improved by the use of a technique called "charge mixing." Although it can be quite effective in *continued on page 4*

SciDAC SPECIAL

A science journal features research on petascale enabling technologies

The current issue of the Cyberinfrastructure Technology Watch (CTWatch) Quarterly included significant contributions — five out of the nine articles — from CRD researchers, who authored papers on code performance, software tools, visualization, scientific data management and data placement solutions for distributed petascale science.

The issue, published in November, featured articles written by dozens of researchers about SciDAC's Centers for Enabling Technologies. SciDAC (Scientific *continued on page 2*

SciDAC *continued from page 1*

Discovery through Advanced Computing), run by the DOE Office of Science, funds projects that investigate obstacles for carrying out petascale computing and develop hardware and software tools to solve them.

The quarterly is published by CTWatch, an online venue for science and engineering community on cyberinfrastructure technology. CTWatch is supported by the National Science Foundation's Cyberinfrastructure Partnership between the San Diego Supercomputing Center (SDSC) and the National Center for Supercomputing Applications (NCSA). SDSC and NCSA run the forum along with the Innovative Computing Laboratory at the University of Tennessee.

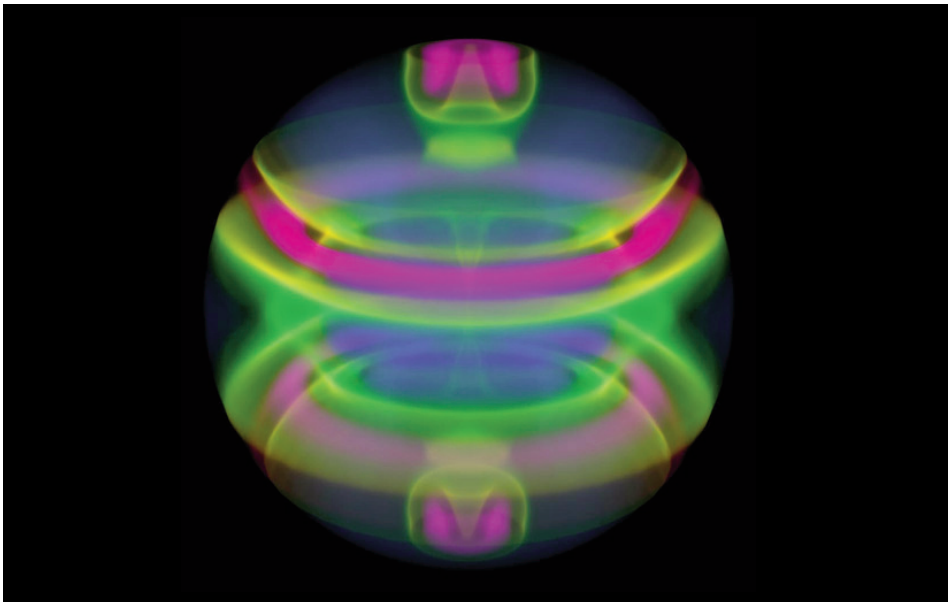
Nine CRD researchers contributed to the current CTWatch Quarterly. David Bailey, chief technologist in CRD, was lead author for the article "Performance Engineering: Understanding and Improving the Performance of Large-Scale Codes." The article identified issues that hamper code performance and ways to help scientists improve their codes and get more out of supercomputers: performance modeling of applications and systems, automatic performance tuning, and application engagement and tuning.

Bailey and his fellow researchers in SciDAC's Performance Engineering Research Institute noted that researchers often devote more programming efforts on analyzing results instead of improving code performance. Yet improving code performance could lead to huge savings. "A quick calculation shows that if one can increase by just 30 percent the performance of two of the major SciDAC applications codes (which together use, say, 10 percent of the NERSC and ORNL high-end systems over three years), this represents a savings of some $6 million," the article said.

Kathy Yelick, head of CRD's Future Technologies Group and incoming director of the National Energy Research Scientific Computing Center, as well as Dan Gunter in the Cyber Infrastructure Development group co-authored the article.

Yelick also contributed to the article "Creating Software Tools and Libraries for Leadership Computing."

Wes Bethel, head of CRD's Visualization group, authored the piece "DOE's SciDAC Visualization and Analytics Center for Enabling Technologies (VACET) – Strategies for Petascale Visual Data Analysis Success." The article outlined the role and techniques of visualization in transforming large sets of observed or simulated data into images that explain the scientific discoveries. Bethel, who is co-principal investigator for the VACET project, gave examples of the types of research that have benefited from using visualization tools, such as astrophysics, combustion and particle physics.



Two types of features are immediately visible in this image showing the entropy field of a radiation/hydrodynamic simulation that models the accretion-induced collapse of a star, a phenomena that produces supernovae. One feature is the sandwiching of high values of entropy between lower values. The other is an overall sense of 3D structure. (Simulation data courtesy of Adam Burrows, University of Arizona, SciDAC Science Application "The Computational Astrophysics Consortium," image courtesy of the Visualization Group, Lawrence Berkeley National Laboratory.)

"As a Center for Enabling Technology, VACET's mission is the creation of usable, production-quality visualization and knowledge discovery software infrastructure that runs on large, parallel computer systems at DOE's Open Computing facilities, and that provides solutions to challenging visual data exploration and knowledge discovery needs of modern science, particularly the DOE science community," the article said.

Several researchers from Bethel's group contributed to the article: Cecilia Aragon, Prabhat and Gunther Weber.

Arie Shoshani, head of SciDAC's Scientific Data Management Center, was lead author of the article called "Scientific Data Management: Essential Technology for Accelerating Scientific Discoveries." The piece discussed the software tools the center has developed for managing and analyzing large quantities of data, from organizing files to getting search results promptly.

One of the search tools highlighted by Shoshani was FastBit, which uses a compressed bitmap to present indexed data. "FastBit is 12 times faster than any known compressed bitmap index in answering range queries," Shoshani wrote. "Because of its speed, FastBit facilitates real-time analysis of data, searching over billions of data values in seconds."

Finally, Dan Gunter and Brian Tierney were co-authors of the article "End-to-End Data Solutions for Distributed Petascale Science." The article described the work by SciDAC's Center for Enabling Distributed Petascale Science, which focuses on developing software tools for quickly transferring and retrieving data, as well as monitoring and troubleshooting any issues during the process.

Read the latest CTWatch Quarterly at http://www.ctwatch.org/quarterly.

Nano Letters *continued from page 1*

ZnO/ZnS and ZnO/ZnTe Heterostructures for Photovoltaic Applications," ranked No. 10 and was also co-authored by Paul Alivisatos, head of the Material Science Division at Berkeley Lab.

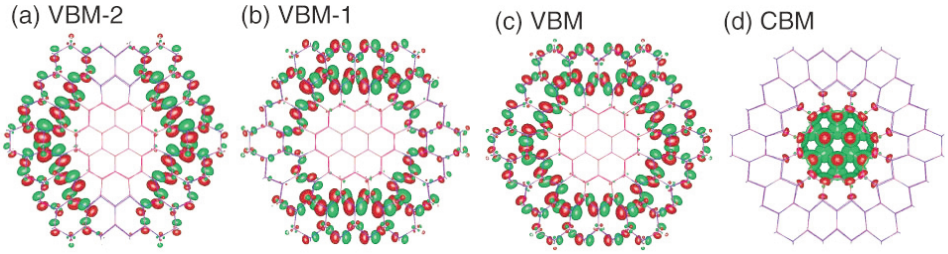
The high cost of the conventional photovoltaic material, silicon, and the scant alternatives for building solar energy equipment for the mass market have prompted scientists to explore nanostructure devices. Previous experimental work at Berkeley Lab by Alivisatos and coworkers has demonstrated how semiconductor nanocrystals can be used to make inexpensive devices. However, these nanocrystals have typically been made from materials such as cadmium selenide (CdSe) and cadmium telluride (CdTe), which contain expensive and toxic elements.

Schrier and his collaborators turned to zinc oxide (ZnO), an abundant, non-toxic and stable compound. ZnO alone isn't an ideal material because it has a large band gap of 3.4 electron volt (eV). But stacked with another material, it could create a composite structure with a much smaller band gap, the researchers wrote, and they tested their theory with zinc sulfide (ZnS) and zinc telluride (ZnTe).

Using resources at the National Energy Research Scientific Computing (NERSC) center, the scientists calculated the optical properties of ZnO/ZnS and ZnO/ZnTe using plane-wave norm-conserving pseudopotential density function theory (DFT) with the PETot code. They include a band-corrected pseudopotential scheme because DFT is inadequate for calculating the value of the band gap.

They studied the composite materials in two forms: planar superlattices and core/shell quantum wires. The researchers considered a superlattice model composed of ZnO/ZnS and another composed of ZnO/ZnTe. They also examined whether applying strain to either superlattice model could further narrow its band gap. With the quantum wires, the scientists only considered the structure composed of ZnO/ZnS.

Results of their work demonstrated significant reductions of the band gap for using both composite materials and structures, compared with using bulk ZnO alone. Strained superlattices showed a further reduction.



Band-edge wave functions of the ZnO/ZnS core/shell nanowire, with positive and negative signs indicated by green and red, respectively. VBM is valence band maximum. CBM is conduction band minimum.

Moreover, the ZnO/ZnS core/shell nanowires proved to be a better structure than superlattices for potential use in photovoltaic devices. "The ZnO/ZnS core/shell nanowire improves upon both the band gap and oscillator strength of its superlattice counterpart," the researchers noted in the paper.

Using the Shockley-Quiesser model for calculating the idea solar cell efficiency, Schrier and his fellow researchers obtained an efficiency limit of 19 percent for the ZnO/ZnS superlattice, 30 percent for the ZnO/ZnTe superlattice and 23 percent for the ZnO/ZnS core/shell quantum wires. The bulk ZnO alone, on the other hand, could only yield a limit of 7 percent.

You can see the list and read the paper at http://pubs.acs.org/journals/nalefd/promo/most/most_accessed/index.html.

Hall of Fame

Kudos for Visualization Poster

A group of CRD researchers won the Best Poster Award at IEEE VAST 2007 (IEEE Symposium on Visual Analytics Science and Technology), which took place from Oct. 30 to Nov. 1 in Sacramento.

Cecilia Aragon, Stephen Bailey, Sarah Poon, Karl Runge and Rollin Thomas were recognized for their poster "Sunfall: A Collaborative Visual Analytics System for Astro-physics," which described the first visual analytics system in production use at a major astrophysics project (the Nearby Supernova Factory).

Cecilia Aragon

Aragon is a member of CRD's Visualization Group and the NERSC Analytics Team. Bailey, Poon, Runge and Thomas were all with the Physics Division when the research was performed (Bailey and Poon have since left the Lab). Thomas has since joined the Computational Cosmology Center (C3), which includes researchers from the CRD and Physics Divisions.

A two-page abstract of the poster can be read at http://vis.lbl.gov/Publications/2007/Sunfall_VAST07.pdf, and the poster itself can be seen at http://vis.lbl.gov/Publications/2007/Sunfall_VAST07_poster.pdf.

SIAM Editorial Board

Esmond Ng, head of CRD's Scientific Computing Group, has accepted an invitation to serve on the editorial board of the SIAM Journal on Scientific Computing. He will begin

continued on page 5

IMG/M *continued from page 1*

from specimens cultured in labs.

IMG/M provides tools for conducting comparative analysis of metagenome datasets integrated with a comprehensive collection of reference isolate microbial genome datasets. For the research on microbes residing in termite guts, Ernest Szeto from CRD's Biological Data Management and Technology Center (BDMTC) provided critical support for maintaining multiple versions of the dataset in IMG/M and for extending data analysis tools in response to the study's needs.

"IMG/M has proven to be an extremely useful resource and tool for analyzing our metagenomic data," said Jared R. Leadbetter, associate professor of environmental microbiology at the California Institute of Technology, and collaborator on the termite hindgut microbial community for bioenergy project. "Such datasets are large, complex and potentially unwieldy. Importantly, IMG/M is more than just an excellent tool to analyze data. The manner in which the results of that analysis are organized and made accessible through a user-friendly interface allows the researcher to rapidly move in a number of different intellectual directions. As a result, the user becomes better educated with and gets a real 'feel' for the data in a manner that would not otherwise be possible on such short timescales."

The research, "Metagenomic and Functional Analysis of Hindgut Microbiota of a Wood-Feeding Higher Termite," appeared in the Nov. 22 issue of Nature. Lead author Falk Warnecke and members of the research team traveled to Costa Rica in search of termites and the microbes that hold the secret to efficient energy conversion. From a colony of termites in the genus Nasutitermes, the researchers extracted the stomach content of 165 specimens and purified the genetic materials for sequencing at the JGI.

The sequencing work yielded 71 million letters of fragmented genetic code, which provides valuable information on the identities of the microbes and the enzymes that they produce. After assembling and analyzing them using the IMG/M, the scientists identified two major bacterial lineages, treponemes and fibrobacters. Researchers have long known the existence of treponemes. But fibrobacters, as it turned out, were a new find.

"The dataset provided by Warnecke et al. is a treasure trove for researchers," wrote Andrea Brune, a researcher in the Department of Biochemistry at the Max Planck Institute for Terrestrial Microbiology in Germany, in a separate article in the same issue of *Nature*.

The researchers identified more than 500 genes related to the enzymatic deconstruction of cellulose and hemicellulose, building blocks of wood.

"Termites can efficiently convert milligrams of lignocellulose into fermentable sugars in their tiny bioreactor hindguts. Scaling up this process so that biomass factories can produce biofuels more efficiently and economically is another story," said Eddy Rubin, JGI Director. "To get there, we must define the set of genes with key functional attributes for the breakdown, and this study represents an essential step along this path."

The termite gut metagenome data will become publicly available in IMG/M in January 2008. More information about IMG/M can be found at http://img.jgi.doe.gov/cgi-bin/m/main.cgi.

The paper in *Nature* can be found at http://www.nature.com/nature/journal/v450/n7169/full/nature06269.html.

David Gilbert at the JGI contributed to the story.

Genome Analysis Workshops

The Biological Data Management and Technology Center (BDMTC) will take part in three microbial genome analysis workshops next year that will provide training in microbial genome and metagenome analysis.

The Microbial Genomics and Metagenomics Workshop will take place at the DOE Joint Genome Institute (JGI). The first five-day workshop will occur on January 7-11, 2008. All three workshops will offer tutorials on a variety of data analysis tools, in particular the tools provided by the Integrated Microbial Genomes (IMG) family of system developed by BDMTC in collaboration with the Genome Biology Program (GBP) at the JGI. Nikos Kyrpides, head of GBP, David Gilbert, head of public relations for the JGI, and Victor Markowitz, head of the BDMTC, are the workshop organizers.

Markowitz also is featured to speak about the IMG-Expert Review (IMG-ER) and IMG-Educational (IMG-EDU), two recently released systems from the IMG family. Krishna Palaniappan of BDMTC will speak about the challenges of integrating different types of genomic data for comparative analyses.

Find out more about the workshop at http://www.jgi.doe.gov/meetings/mgm.

DCM Algorithm

continued from page 1

practice, a charge mixing-enabled SCF iteration may not reduce the total energy approximation monotonically, and there is no guarantee that charge mixing will always fix the convergence failure of the SCF iteration.

The DCM algorithm, on the other hand, is designed to minimize the total energy of the atomistic system directly. Through the use of adaptive "trust regions" of appropriate sizes, the reduction of the total energy is guaranteed to be monotonic.

Chao Yang, a member of the Scientific Computing Group, was the lead author. Juan Meza, head of the High Performance Computing Research Department, and Lin-Wang Wang in the Scientific Computing Group co-authored the paper.

The researchers viewed the SCF iteration as an indirect way of minimizing the total energy function through the minimization of a sequence of quadratic surrogate functions. They noticed that, at each SCF iteration, the surrogate function shares the same gradient with that of the total energy function at the current approximation. Hence moving along a descent direction associated with the surrogate function near the current approximation is likely to result in a reduction in the total energy. However, if one moves too far along that direction, the total energy may actually increase, resulting in convergence failure of the SCF iteration.

So Yang and his fellow researchers proposed to stabilize the convergence of the SCF iteration by imposing an additional constraint *continued on page 6*

Hall of Fame continued from page 3

Esmond Ng

the three-year term on January 1.

SIAM, the Society for Industrial and Applied Mathematics, publishes 14 journals. Ng, whose research interests include sparse matrix computation, numerical linear algebra and parallel computing, also is on the editorial board of the SIAM Journal on Matrix Analysis and Applications. He has been on the board since 1997 and is now serving the fourth term.

Nano Poster Wins Award

Three researchers from won the Best Poster Award at the November SC07 supercomputing conference in Reno.

Zhengji Zhao, Juan Meza and Lin-Wang Wang were recognized for their poster describing "A New O(N) Method for Petascale Nanoscience Simulations," which describes new linear scaling three-dimensional fragment (LS3DF) method for ab initio electronic structure calculations.

The poster was one of 39 accepted for the conference from more than 150 submissions. SC07 is the leading international conference on high performance computing, networking, storage and analysis.

Zhao is a high-performance computing consultant at the National Energy Research Scientific Computing (NERSC) center at Berkeley Lab. Meza is the head of CRD's High Performance Computing Research Department and Wang is a scientist in the department's Scientific Computing Group.

Serving SIAG/SC

Ali Pinar, a researcher in the Scientific Computing Group, has been elected to serve as secretary for the SIAM Activity Group on Supercomputing (SIAG/SC). His two-year term begins on January 1.

Pinar's research focuses on combinatorial scientific computing, and he is particularly interested in tackling combinatorial problems that are directly associated with scientific and engineering goals. He has worked on vulnerability analysis of the electric power grid, interconnection networks for ultra-scale systems, energy efficient disk systems, and supernova spectra analyses.

SIAM is the Society for Industrial and Applied Mathematics. The SIAM Activity Group on Supercomputing provides a forum for computational mathematicians, computer architects and computational scientists to exchange ideas on mathematical algorithms and computer architecture needed for high-performance computer systems. The activity group promotes the exchange of ideas by focusing on the interplay of analytical methods, numerical analysis and efficient computation.

(From left to right) Lin-Wang Wang, Juan Meza and Zhengji Zhao accepted the Best Poster Award at SC07 from Chuck Koelbel of Rice University.

Promoting Computational Nanoscience

Juan Meza and Kathy Yelick spoke at a November workshop on "Excellence in Computer Simulation" in Berkeley. The workshop is organized and sponsored by the Network for Computational Nanotechnology (NCN), the Center of Integrated Nanomechanical Systems (COINS) and Berkeley Lab's Molecular Foundry.

The meeting created a forum for sharing thoughts about where computational science is heading and ideas on how the nanoscience community can be more effective in its research. It's also an opportunity for students to think about how to prepare themselves for careers in computational science and engineering.

Meza is the head of CRD's High Performance Computing Research Department while Yelick is the incoming director for the National Energy Research Scientific Computing (NERSC) center.

New AAAS Fellow

David Patterson, a CRD researcher and the E.H. and M.E. Pardee Chair of Computer Science at UC Berkeley, has been named a 2007 Fellow of the American Association for the Advancement of Science (AAAS). New fellows will be recognized for their contributions to science and technology at the Fellows Forum on February 16 during the AAAS Annual Meeting in Boston, where each fellow will receive a certificate and a blue and gold rosette as a symbol of their distinguished accomplishments.

continued on page 6

DCM Algorithm

continued from page 4

to the surrogate minimization problem. This additional constraint sets up a region in which the minimizer of the surrogate function can be "trusted," in the sense that a reduction in the surrogate in that region will lead to a reduction in the total energy as well.

However, identifying the appropriate size of a trust region would require computing all eigenvalues of a fixed Kohn-Sham Hamiltonian, which is not practical due to the dimension of the Hamiltonian. If the size of the trust region is too large, the SCF iteration may fail to converge. If the trust region is too small, the SCF iteration may converge very slowly.

To overcome this difficulty, the researchers came up with an alternative approach that minimizes the total energy directly. In this direct constrained minimization (DCM) algorithm, the total energy functional is projected into a sequence of overlapping subspaces. The projected problem has a significantly fewer degrees of freedom, thus can be solved efficiently by the trust region-enabled SCF iteration.

The optimal solution of the projected problem yields an optimal search direction and step length for updating the approximation to the minimizer of the true total energy function. They demonstrated this method using two numerical examples, which showed that "such a scheme outperforms the SCF iteration combined with charge mixing in terms of both efficiency and reliability."

Read about the research at SIAM *Journal of Scientific Computing.*

About CRD Report

CRD Report, which publishes every other month, highlights the cutting-edge research conducted by staff scientists in areas including turbulent combustion, nano materials, climate change, distributed computing, high-speed networks, astrophysics, biological data management and visualization. CRD Report Editor Ucilia Wang can be reached at 510 495-2402 or Uwang@lbl.gov. Find previous CRD Report articles at http://crd.lbl.gov/html/news/CRDreport.html.

Hall of Fame

continued from page 5

Patterson is also a member of the National Academy of Sciences and the National Academy of Engineering. He led the design and implementation of RISC I, the first VLSI Reduced Instruction Set Computer, and he shared the 2000 IEEE von Neumann medal and IEEE's 1999 Reynold Johnson Information Storage Award. He is currently building novel microprocessors using intelligent DRAM for portable multimedia devices, and is a member of the ROC (Recovery Oriented Computing) project.

New Staff

The Scientific Computing Group welcomes Dr. Nenad Vukmirovic, who has joined them as a postdoctoral fellow. He will work with Lin-Wang Wang on electronic structure calculations for large organic molecules, with a focus on the charge patching method.

Nenad Vukmirovic

Nenad received his Ph.D. from the University of Leeds in the United Kingdom this past summer. In his Ph.D. work, Nenad simulated electron devices, lasers, quantum dots and electron transport. He also has experience in developing large-scale scientific software. During his graduate studies, Nenad won several IEEE fellowships. As a high school student, Nenad was a gold medalist at the 30th International Physics Olympiad in 1999.

Sandra Wittenbrock has joined the CRD/High Performance Computing Research Department System Engineering Team. She has been working at Berkeley Lab since 2002. She will provide computer systems support for the division, working with users and staff/scientists to provide a stable computer and infrastructure environment. She will also work on long-term projects to promote better infrastructure support.

Horst Simon

Supercomputer in South Africa

Forbes.com quoted Horst Simon, Associate Lab Director for Computing Sciences at Berkeley Lab, in a recent article about IBM's donation of a BlueGene/P to the Center for High Performance Computing in Cape Town, South Africa.

The story looked at what the supercomputer can do to spur scientific discoveries and technical innovations in South Africa and the continent. The supercomputer is five times more powerful than the next fastest computer on the continent, according to Forbes.com.

Simon pointed out that one supercomputer alone isn't enough to tackle key issues such as modeling climate change, predicting the spread of infectious diseases and developing more efficient mining operations.

Simon added that IBM's donation would work well in tandem with other programs such as the One Laptop Per Child by Nicholas Negroponte. "The OLPC program builds computer literacy from the ground up and gets a large number of mathematically inclined individuals involved," Simon said in the article. "In fact, we need both of these approaches."

Read more about the IBM donation and Simon's view at Forbes.com.