DISTRIBUTED COMPUTING

NEWS

Grassroots Supercomputing

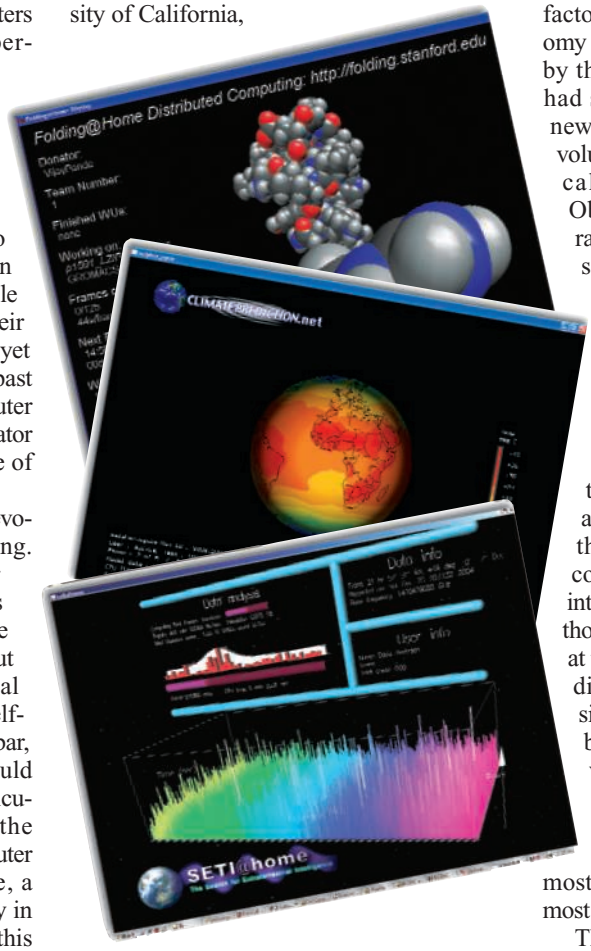
What started out as a way for SETI to plow through its piles of radio-signal data from deep space has turned into a powerful research tool as computer users across the globe donate their screen-saver time to projects as diverse as climate-change prediction, gravitational-wave searches, and protein folding

OXFORD, U.K.—If Myles Allen and David Stainforth had asked for a supercomputer to test their ideas about climate change, they would have been laughed at. In order to push the limits of currently accepted climate models, they wanted to simulate 45 years of global climate while tweaking 21 parameters at once. It would have required a supercomputer's fully dedicated attention over years, preempting the jealously guarded time slots doled out to many other projects. "Doing this kind of experiment wasn't even being considered," recalls Stainforth, a computer scientist here at Oxford University. So instead, he and Oxford statistician Allen turned to the Internet, where 100,000 people from 150 countries donated the use of their own computers—for free. Although not yet as flexible, their combined effort over the past 2 years created the equivalent of a computer about twice as powerful as the Earth Simulator supercomputer in Yokohama, Japan, one of the world's fastest.

Stainforth's project is part of a quiet revolution under way in scientific computing. With data sets and models growing ever larger and more complex, supercomputers are looking less super. But since the late 1990s, researchers have been reaching out to the public to help them tackle colossal computing problems. And through the selfless interest of millions of people (see sidebar, p. 812), it's working. "There simply would not be any other way to perform these calculations, even if we were given all of the National Science Foundation's supercomputer centers combined," says Vijay Pande, a chemical biologist at Stanford University in Palo Alto, California. The first fruits of this revolution are just starting to appear.

World supercomputer

Strangely enough, the mass participation of the public in scientific computing began with a project that some scientists believe will never achieve its goal. In 1994, inspired by the 25th anniversary of the moon landing, software designer David Gedye wondered "whether we would ever again see such a singular and positive event," in which people across the world join in wonder. Perhaps the only thing that could have that effect, thought Gedye, now based in Seattle, Washington, would be the discovery of extraterrestrial intelligence. And after teaming up with David Anderson, his former computer science professor at the University of California,



Strength in numbers. Millions of computers now crunch data for diverse research projects.

Berkeley, and Woody Sullivan, a science historian at the University of Washington, Seattle, he had an idea how to work toward such an event: Call on the public to get involved with the ongoing Search for Extraterrestrial Intelligence (SETI) project.

In a nutshell, SETI enthusiasts argue that we have nothing to lose and everything to gain by scanning electromagnetic radiation such as radio waves—the most efficient method of interstellar communication we know of—from around the galaxy to see if anyone out there is broadcasting. After the idea for SETI was born in 1959, the limiting factor at first was convincing radio astronomy observatories to donate their help. But by the mid-1990s, several SETI projects had secured observing time, heralding a new problem: how to deal with the huge volume of data. One Berkeley SETI project, called SERENDIP, uses the Arecibo Observatory in Puerto Rico, the largest radio telescope in the world, to passively scan the sky around the clock, listening to 168 million radio frequencies at once. Analyzing this data would require full-time use of the Yokohama Earth Simulator, working at its top speed of 35 teraFLOPS (10^{12} calculations per second).

Gedye and his friends approached the director of SERENDIP, Berkeley astronomer Daniel Werthimer, and posed this idea: Instead of using one supercomputer, why not break the problem down into millions of small tasks and then solve those on a million small computers running at the same time? This approach, known as distributed computing, had been around since the early 1980s, but most efforts had been limited to a few hundred machines within a single university. Why not expand this to include the millions of personal computers (PCs) connected to the Internet? The average PC spends most of its time idle, and even when in use most of its computing power goes untapped.

The idea of exploiting spare capacity on PCs was not a new one. Fueled by friendly competition among hackers, as well as cash prizes from a computer security company, thousands of people were already using their PCs to help solve mathematical problems. A trailblazer among these efforts was GIMPS, the Great Internet Mersenne Prime Search, named after the 16th century French monk who discovered a special class of enormous numbers that take the form $2^P - 1$ (where P is a prime). GIMPS founder George Woltman, a programmer in Florida, and Scott Kurowski,

CREDITS (TOP TO BOTTOM): PHOTOS.COM; FOLDING@HOME; CLIMATEPREDICTION.NET; SETI@HOME

Published by AAAS

a programmer in California, automated the process and put a freely downloadable program on the Internet. The program allowed PCs to receive a task from the GIMPS server, "crunch" on it in the background, and send the results back without the PC user even noticing.

Using computer time in this way is not always a blameless activity. In 1999, system administrator David McOwen marshaled hundreds of computers at DeKalb Technical College in Clarkston, Georgia, to crunch prime numbers with a program from a distributed network—but without getting permission. When found out, he was arrested and accused of costing the college more than $400,000 in lost bandwidth time. But the case never came to court, and McOwen accepted penalties of 80 hours of community service and a $2100 fine. The previous year, computer consultant Aaron Blosser got the computers of an entire Colorado phone company busy with GIMPS. Because his supervisor had given him permission to do so, he was not charged, but because at the time it was considered a potential act of Internet terrorism, the FBI confiscated his computers.

Undaunted, Gedye and his team set about carving up the SETI processing work into bite-sized chunks, and in 1999 the team went public with a screen-saver program called SETI@home. As soon as a PC went idle, the program went to work on 100-second segments of Arecibo radio data automatically downloaded from the Internet, while the screen saver showed images of the signal analysis. It took off like wildfire. Within 1 month, SETI@home was running on 200,000 PCs. By 2001, it had spread to 1 million. Public-resource computing, as Anderson calls it, was born.

So far at least, SETI@home hasn't found an ET signal, admits Anderson, and the portion of the galaxy searched "is very, very limited." But the project has already accomplished a great deal: It not only fired up the public imagination, but it also inspired scientists in other fields to turn to the public for help tackling their own computing superproblems.

Democratizing science?

Stanford's Pande, who models how proteins fold, was among the first scientists to ride the public-resource computing wave. Proteins are like self-assembling puzzles for which we know all the pieces (the sequence of amino acids in the protein backbone) as well as the final picture (their shape when fully folded), but not what happens in between. It only takes microseconds for a typical protein to fold itself up, but figuring out how it does it is a computing nightmare. Simulating nanosecond slices of folding for a medium-sized protein requires an entire day of calculation on the fastest machines and years to finish the job. Breaking through what Pande calls "the microsecond barrier" would not only help us understand the physical chemistry of normal proteins, but it could also shed light on the many diseases caused by misfolding, such as Alzheimer's, Parkinson's, and Creutzfeldt-Jakob disease.

A year after SETI@home's debut, Pande's research group released a program called Folding@home. After developing new methods to break the problem down into workable chunks, they crossed their fingers, hoping that enough people would take part. For statistical robustness, identical models with slightly tweaked parameters were doled out in parallel to several different PCs at once, so success hinged on mass participation.

The simulations flooded back. By the end of its first year, Folding@home had run on 20,000 PCs, the equivalent of 5 million days of calculation. And the effort soon proved its worth. Pande's group used Folding@home to simulate how BBA5, a small protein, would fold into shape starting only from the protein's sequence and the laws of physics. A team led by Martin Gruebele, a biochemist at the University of Illinois, Urbana-Champaign, tested it by comparing with real BBA5. The results, reported in 2002 in *Nature*, showed that Folding@home got it right. This marks "the first time such a convergence between theory and experiment could be made," says Pande.

Public-resource computing now has the feel of a gold rush, with scientists of every stripe prospecting for the bonanza of idle computing time (see table, left). Biological projects dominate so far, with some offering screen savers to help study diseases from AIDS to cancer, or predict the distribution of species on Earth. But other fields are staking their own claims. Three observatories in the United States and Germany trying to detect the fleeting gravitational waves from cataclysmic events in space—a prediction of Einstein's—are doling out their data for public crunching through a screen saver called Einstein@home. Meanwhile, CERN, the European particle physics laboratory near Geneva, Switzerland, is tapping the public to help design a new particle accelerator, the Large Hadron Collider. LHC@home simulates the paths of particles whipping through its bowels.

Project/URL	Research Base	Goal
Mersenne Prime Search www.mersenne.org	Worldwide	Identify enormous prime numbers
SETI@home setiathome.ssl.berkeley.edu	UC Berkeley	Find extraterrestrial intelligence
Folding@home folding.stanford.edu	Stanford	Predict how proteins fold
ClimatePrediction.net climateprediction.net	Oxford	Test models of climate change
LHC@home lhcathome.cern.ch	CERN	Model particle orbits in accelerator
Einstein@home einstein.phys.uwm.edu	U.S. and Germany	Identify gravitational waves
Cancer Research Project www.grid.org/projects/cancer	NCI and Oxford	Search for candidate drugs against cancer
Lifemapper www.lifemapper.org	University of Kansas	Map global distribution of species

The projects launched so far have only scraped the surface of available capacity: Less than 1% of the roughly 300 million idle PCs connected to the Internet have been tapped. But there are limits to public-resource computing that make it impractical for some research. For a project to make good use of the free computing, says Stainforth, "it has to be sexy and crunchable." The first factor is important for attracting PC owners and persuading them to participate. But the second factor is "absolutely limiting," he says, because not all computational problems can be broken down into small tasks for thousands of independent PCs. "We may have been lucky to have chosen a model that can be run on a typical PC at all," Stainforth adds.

In spite of those limitations, the size and number of public-resource computing projects is growing rapidly. Much of this is thanks to software that Anderson developed and released last year, called Berkeley Open Infrastructure for Network Computing (BOINC). Rather than spending time and money developing their own software, researchers can now use BOINC as a universal template for handling the flow of data. In a single stroke, says Anderson, "this has slashed the cost of creating a public-resource computing project from several hundreds of thousands of dollars to a few tens of thousands." Plus, BOINC vastly improves the efficiency of the entire community by allowing PCs to serve several research projects at once: When one project needs a breather, another can swoop in rather than leaving the PC idle.

Published by AAAS

It works, too

As the data streams in from the many projects running simultaneously on this virtual supercomputer, some researchers are getting surprising results. To the initial dismay of CERN researchers, LHC@home occasionally produced very different outputs for the same model, depending on what kind of PC it ran on. But they soon discovered that it was caused by "an unexpected mathematical problem," says François Grey, a physicist at CERN: the lack of international standards for handling rounding errors in functions such as exponential and tangent. Although the differences between PCs were minuscule, they were amplified by the sensitive models of chaotic particle orbits. The glitch was fixed by incorporating new standards for such functions into the program.

The results of ClimatePrediction.net have been surprising for a different reason. "No one has found fault with the way our simulations were done," says Stainforth. Instead, climate scientists are shocked by the predictions. Reporting last January in *Nature*, a team led by Stainforth and Allen found versions of the currently accepted climate model that predict a much wider range of global warming than was thought. Rather than the consensus of a 1.5° to 4.5°C increase in response to a doubling of atmospheric CO_2, some simulations run on the Oxford screen saver predict an 11°C increase, which would be catastrophic. Critics argue that such warming is unrealistic because the paleoclimate record has never revealed anything so dramatic, even in response to the

Grid Sport: Competitive Crunching

You won't find the names of Jens Seitler, Honza Cholt, John Keck, or Chris Randles among the authors of scientific papers. Nor, for that matter, the names of any of the millions of other people involved with the colossal computing projects that are predicting climate change, simulating how proteins fold, and analyzing cosmic radio data. But without their uncredited help, these projects would be nonstarters.

In the 6 years since the SETI@home screen-saver program first appeared, scientists have launched dozens of Internet projects that rely on ordinary people's computers to crunch the data while they sit idle. The result is a virtual computer that dwarfs the top supercomputer in speed and memory by orders of magnitude. The price tag? Nothing. So who are these computer philanthropists? The majority seem to be people who hear about a particular project that piques their interest, download the software, and let it run out of a sense of altruism. Others may not even be aware they are doing it. "I help about a dozen friends with repairs and upgrades to their PCs," says Christian Diepold, an English literature student from Germany, "and I install the [screen-saver software] as a kind of payment. Sometimes they don't even know it's on there."

But roughly half of the data processing contributed to these science projects comes from an entirely different sort of volunteer. They call themselves "crunchers," and they get kicks from trying to churn through more data than anyone else. As soon as the projects began publishing data-crunching statistics, competition was inevitable. Teams and rank ladders formed, and per capita crunching has skyrocketed. "I'm addicted to the stats," admits Michael, a member of a cruncher team called Rebel Alliance. To get a sense of what makes them tick, *Science* interviewed dozens of crunchers in the Internet chat forums where they socialize.

Team players. Honza Cholt says crunchers have deep discussions about the science.

Interest in crunching does not appear to correlate strongly with background. For their day jobs, hard-core crunchers are parking lot attendants, chemical engineers, stay-at-home moms and dads, insurance consultants, and even, in at least one case, miners. Their distribution, like the Internet, is global. What's the motive? People crunch "for a diversity of reasons," says Randles, a British accountant who moderates the forum for ClimatePrediction.net, but altruism tops the list. "After losing six friends over the last 2 years to cancer, I jumped at the chance to help," says an electrician in Virginia who goes by the username JTWill and runs the Find-a-Drug program on his five PCs. As a systems administrator named Josh puts it, "Why let a computer sit idle and waste electricity when you could be contributing to a greater cause?"

But another driving force is the blatant competition. Michael of Rebel Alliance has recently built a computer from scratch for the sole purpose of full-time crunching, but he says he still can't keep up with Stephen, a systems engineer in Missouri and self-proclaimed "stats junkie" who crunches on 150 computers at once. Without the competition, "it wouldn't be as much fun," says Tim, a member of Team Anandtech who crunches for Folding@home. And like any sport, rivalries are soon simmering. "Members from different teams drop in on each other's forums and taunt each other a bit," says Andy Jones, a cruncher in Northern Ireland, "but it's all in good humor." As Anandtech team member Wiz puts it, "What we have here is community."

But where does this leave the science? Do crunchers care how the fruits of their labor are used, or do they leave it all to the researchers? It depends on the project, says Cholt, a sociology student in the Czech Republic, "but the communities that form often have long and deep discussions about the science." What holds the core of the crunching community together, says Seitler, a computer specialist in Germany, is the chance "for normal people to take part in a multitude of scientific projects." In some cases, crunchers have even challenged the researchers' published conclusions. "Many scientists would groan at the thought of nonscience graduates questioning their work," says Randles, but "scrutiny beyond peer review seems an important aspect to science."

Far from indifferent, crunchers can become virtual members of the research team, says François Grey, a physicist at CERN, the particle physics lab near Geneva, Switzerland, who helps run LHC@home. Above and beyond donating their computers, "they actually help us solve problems and debug software. And you have to keep them informed about what's going on with the project, or they get upset." Crunchers might not get credited on papers, says Grey, but "scientists have to treat this community with respect."

–J.B.

CREDITS: HONZA CHOLT

Published by AAAS

largest volcanic eruptions. Stainforth emphasizes that his method does not yet allow him to attach probabilities to the different outcomes. But the upshot, he says, is that "we can't say what level of atmospheric carbon dioxide is safe." The finding runs against recent efforts to do so by politicians.

And according to Stainforth, this illustrates something that makes public-resource computing a special asset to science. Rather than a hurdle to be overcome, "public participation is half of the goal." This is particularly true for a field like climate prediction, in which the public can influence the very system being studied, but it may also be true for less political topics. "We in the SETI community have always felt that we were doing the search not just for ourselves but on behalf of all people," says Sullivan. What better way to "democratize" science than to have a research group of several million people?

–JOHN BOHANNON

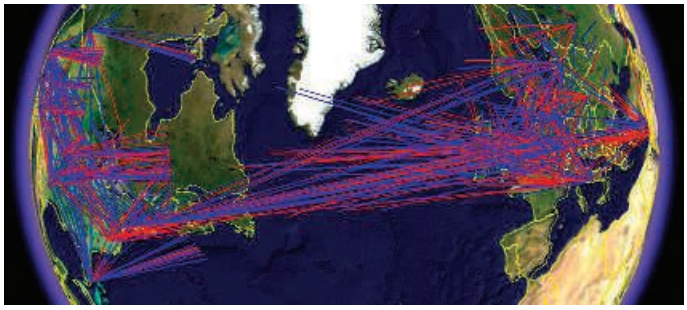
John Bohannon is a science writer based in Berlin.

NEWS

Data-Bots Chart the Internet

It's hard to map the global Internet from a small number of viewpoints. The solution may be to enlist computer users worldwide as local cartographers of cyberspace

Anyone who has tried to study the twists and turns in the data superhighway knows the problem: It is difficult even to get a decent map of the Internet. Because it grew up in a haphazard fashion with no structure imposed, no one knows how the myriad telephone lines and satellite links weave together its more than 300,000,000 computers. Today's best maps offer a badly distorted picture, incomplete and biased by a U.S. viewpoint, hampering computer scientists' efforts to design software that would make the Internet more stable and less prone to attack. But a new mapping effort may succeed where others have failed. "We want to let the Internet measure itself," says computer scientist Yuval Shavitt of Tel Aviv University in Israel, who, along with colleagues, hopes to enlist many thousands of volunteers worldwide to take part in the effort.



Gridlock. Accurate Internet maps could provide users with data traffic reports.

CREDIT: YUVAL SHAVITT AND ERAN SHIR/TEL AVIV UNIVERSITY

At the lowest level, the computers that comprise the Internet are known as "routers." They carry out the basic information housekeeping of the Net, shuttling e-mails and information packets to and fro. At a somewhat higher linked-facility level, however, the Internet can also be viewed as a network of subnetworks, or "autonomous systems," each of which corresponds to an Internet service provider or other collection of routers gathered together under a single administration. But how is this network of networks wired up?

Two years ago, computer scientist Kimberly Claffy and colleagues from the Cooperative Association for Internet Data Analysis at the University of California, San Diego, used a form of Internet "tomography" to find out. They sent out information-gathering packets from 25 computers to probe over 1 million different destinations in the Internet. Along the way, each packet recorded the links along which it moved, thereby tracing out a single path through the Internet—a chain of linked autonomous systems. Putting millions of such paths together, the researchers eventually built up a rough picture of more than 12,000 autonomous systems with more than 35,000 links between them (see www.caida.org/analysis/topology/as_core_network).

Through such efforts, researchers now understand that the Internet has a highly skewed structure, with some autonomous systems playing the role of organizing "hubs" that have far more links than most others. But researchers also know that their very best maps are still seriously incomplete.

The trouble is that all mapping efforts to date have started out from a fairly small number of sites, 50 at the most. So the maps produced tend to be biased by the locations of those sites. From some computer A, for example, researchers can send probing packets out toward computers B and C and thereby learn paths connecting A to B and A to C. But the probes would be unlikely to explore links between B and C, for the same reason that driving from New York to Boston and from New York to Montreal tells one little about the roads between Boston and Montreal. "If you send probes from only a few points, you naturally get a very partial point of view," says physicist Alessandro Vespignani, an expert on Internet topology at Indiana University, Bloomington.

To overcome this problem, Shavitt and colleagues are pioneering a new approach inspired by the idea of distributed computing. Anyone can now download a program from the Web site www.netdimes.org that will help in a global effort to map the Internet. Using no more than a few percent of the host computer's processing power, the program acts as a software agent, sending out probing packets to map local connections in and around the autonomous system in which the computer sits. "What we ask for is not so much processing power but location," says Shavitt. "We hope that the more places we have presence in, the more accurate our maps will be."

Since the project's inception late last year, individuals have downloaded nearly 800 agents that are now working together to map the Internet from 50 nations spread across all the continents. "We've already mapped out about 40,000 links between about 15,000 distinct autonomous systems, and we can already see that the Internet is about 25% denser than it was previously thought to be," says Shavitt. "This is a great project with a very new perspective," says Vespignani, who points out that better maps will help Internet administrators in predicting information bottlenecks and other hot spots.

Shavitt and his colleagues estimate that once they have about 2000 agents operating, it should be possible to get a complete map of the Internet at the autonomous-system level in less than 2 hours. Once they can do that, they hope to provide individual users with local Internet "weather reports." Ultimately, they would like to map the Internet at the level of individual routers—getting a more detailed map of the physical Internet. "We'll need about 20,000 agents distributed uniformly over the globe to get a good map at that level," says Shavitt. Then there'll be no excuse for getting lost in cyberspace.

–MARK BUCHANAN

Mark Buchanan is a writer in Cambridge, U.K.

Published by AAAS