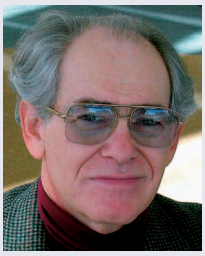
Observatoire Landau



Petascale Simulations May Shed Light on Frailty of Human Condition

By Rubin Landau

In April, the US National Science Foundation (NSF) announced a solicitation for Accelerating Discovery in Science and Engineering through Petascale Simulations and Analysis (PetaApps; NSF 07-559). This competition will fund between 11 and 22 grants of up to US$2 million each. It follows and supports a solicitation for creating a petascale computing environment for science and engineering (NSF 06-573), which is expected to award $200 million in a single grant! Although the deadline for this latter competition has passed (missed your opportunity again?), and the award has yet to be announced, the competition for the $2 million-and-less grants is still open.

In case you need reminding, deca = 10, hector = 10^2, kilo = 10^3, mega = 10^6, giga = 10^9, tera = 10^{12}, peta = 10^{15}, and exa = 10^{18}. It's hard for me to imagine a computer capable of delivering sustained performance of greater than 10^{15} floating-point operations per second on realistic and useful applications involving petabytes of data. Yes, we've witnessed Moore's law on our desktop computers, but at the current rate, it would take 30 years for my desktop to reach the petascale, whereas the NSF proposal envisions such a system by 2011. To be accurate, the IBM BlueGene/L system has already achieved 281 Tflops, so it's not such a big jump to systems employing tens or hundreds of thousand of processors, each containing multiple cores capable of executing multiple threads and, often, arithmetic units that support small vector instructions.

Regardless of the technical achievements required to build a petascale computer, I suspect the reason for the latter NSF solicitation is that it's hard to imagine the type of problems that petascale computing can solve. Even if you could, the simulation, optimization, and analysis tools you would need to solve the problem simply aren't available. Scientists will have to devise algorithms that take advantage of the different types of parallelism available, with multilevel caches, local and main remote memory, intra- and internodal communication networks, parallel I/O, and the associated various levels of latency. We should be aided by the development of the new partitioned global address space compilers that offer simpler programming models such as co-Array Fortran, Unified Parallel C, and Titanium, together with their underlying native-mode communications library. But that's all part of our dreams for the future.

Clearly, you don't use such machines to compile your income taxes. Although we can always look at our old simulations with increasing resolution and accuracy, petascale computing will be most appropriate for new types of science and engineering problems that must be approached in new

also by rich metadata, including user-generated tags. And the human brain is a great source of metadata.

"When you've seen something before, you have very rich memories about it," Dumais says. "If you read an article, you may remember where it was published, what the article looked like, what else you were doing at the time, what people were involved in it. All of that information provides useful retrieval cues." Menczer's study, she points out, showed that searching and browsing are two distinct but complementary ways to access information. Phlat tightly couples both activities: users can search using keywords or metadata, browse through categories of results, and refine their search based on whatever criteria matter to them. One day, people might use a product like Phlat to browse the Web. But there are other improvements Dumais would like to see in search engines, too—ones that would particularly benefit scientists.

"As a researcher, I gather information, but that's only the first step," she says. "I analyze it, contrast it with other things, and then use it in writing a paper or making a presentation or sharing it with colleagues. We are doing a reasonable job of helping people gather information, but I don't think we're doing as good a job at helping people analyze that information. That's how I think we can really help the science and engineering community."

Value Chain

Both Norvig and Dumais say that researchers face unique obstacles when they search for information—the things they want to find, such as journal articles or research data, often aren't available for free on the Web. Christine Borgman, professor and Presidential Chair in Information Studies at the University of California, Los Angeles, knows why those items aren't available. She also adds that researchers have more control over the situation than they know.

Most copyright agreements allow researchers to post their articles online, albeit with some restrictions, but most researchers just don't do it—they don't realize they have the right or they don't want to bother. Either way, they're limiting the number of people who can see and reference their work. And researchers tend not to post their data online because they're afraid they'll lose control over it and people will use the data without crediting its source. If tools were created to make depositing data on the Web easy, while tagging the data to clearly indicate its origins, the situation might change. With open access to journal articles—even drafts and preprints—and data, researchers

ways. As a class, these new problems would involve multitime and space scales, and multiphysics. They include, but aren't limited to, topics such as

- the radiative, dynamic, and nuclear physics of stars and the collision of stars;
- reactions of large biomolecules assemblages, such as cell membranes;
- nonlinear interactions between cloud systems, weather systems, and the climate;
- prediction of 3D protein structures from their primary amino acid sequence;
- determination of the Earth's internal structure via seismic inversions;
- designing molecular electronic devices;
- generation and evolution of magnetic fields in planets and stars;
- galaxy formation and evolution;
- design of specific catalysts, pharmaceuticals, and molecular materials; and
- climate modeling.

No doubt you're asking yourself, "where in this list is the human frailty that was advertised in this column's title?" I think it's us and our upbringing. If pushed, I might be able to imagine petascale simulations, although I suspect that truly imaginative petascale applications will need to come from the next generation of scientists and engineers who were raised to think about problems on this scale. This leads me to the second half of this column. How do we prepare the human resources needed to provide the creativity and originality in petascale computing? The immediate answer is with educational activities focused on trends in high-performance computing. Three such educational programs are:

- 2007 Department of Energy Summer School in Multiscale Mathematics and High Performance Computing (http://multiscale.emsl.pnl.gov/), 29 June–3 July 2007, Oregon State University. The summer program provides introductions to mathematical and computational methods used to model physical systems at various scales, tutorials, instructor-led lab activities, and research talks.
- SC07 Education Program (www.computationalscience.org/workshops/summer07/index.html), 10–13 November 2007, Reno, Nevada. Hands-on activities include how to apply computational science, grid computing, and high-performance computing resources in education.
- TeraGrid '07 (www.union.wisc.edu/teragrid07), 4–8 June 2007, Madison, Wisconsin. The conference's theme is "Broadening Participation in TeraGrid." It features scientific results from the use of TeraGrid and tutorials on TeraGrid resources, such as visualization tools, Science Gateways, and Globus middleware.

See ya there!

could assemble the "value chain" for a particular line of research.

Still, just getting at the scientific information that's already out there isn't easy. Borgman says this is because search engines are oriented toward the naïve searcher, not the scientific searcher. "As long as people are relying on search engines that are getting their money from finding the cheapest deal on airfares and cameras, they're never going to be very useful for science." To be really useful, search engines would have to reveal the existence of data repositories so that researchers could then perform specialized searches inside them. Libraries around the world are expanding their digital archives right now, she points out. If search engines don't reveal these archives, many treasures could go undiscovered.

A New Search Engine

At Indiana University, Menczer is creating a new kind of search engine that's based on a very familiar form of Web metadata—bookmarks. GiveALink (www.givealink.org) is similar to the del.icio.us (http://del.icio.us) social-networking site in that it lets registered users share their bookmarks online and tag them with even more meta information. It's also like the news aggregator Digg (www.digg.com), in that it lets users vote for Web sites they think are important. Visitors upload their bookmarks, and then Menczer's team applies a machine-learning algorithm that maps out relationships between the bookmarks and creates a ranking scheme for the search engine.

For instance, if many people have the same two Web sites stored in the same folder in their bookmarks file, then GiveALink ranks the two sites as closely related. The aggregate data forms a semantic network, or ontology, for the Web. "It's a way to build a classification automatically, by everybody just submitting a piece of the puzzle," Menczer says. "So now I get a more trusted notion of a link, where there is meaning attached to it." Links are weighted to take into account people's notions of how things are related. This means that if any group of people, such as scientists in a particular discipline, donated all their bookmarks, they would build an ontology that was unique to them.

In return for donating their bookmarks to science, users can get personalized recommendations for everything from music to news to podcasts. And Menczer thinks industrial partners could integrate GiveALink's methods into any search engine to help them see relationships among results.