Case Study.

Pioneer is Seeding the Future with High Performance Computing

Compete.

Council on Competitiveness

This material is based on research sponsored by DARPA under agreement number FA8750-07-1-0225. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. government.

Design: Soulellis Studio

Printed in the United States of America

Pioneer is Seeding the Future with High Performance Computing

High performance computing (HPC) is enabling researchers at Pioneer Hi-Bred, a DuPont business, to conduct leading-edge research into plant genetics to create improved varieties of seeds. Pioneer researchers use HPC to manage and analyze massive amounts of molecular, plant, environmental and farm management data, allowing them to make product development decisions much faster than by using traditional experiments and testing alone. HPC also gives Pioneer a window into the future, allowing them to make more informed decisions about their applied breeding programs and frontload experiments with predicted potential winners. For Pioneer, the result has been faster improvement in new seed products, staying ahead of the competition, a major jump in innovation and productivity, and the ability to help meet some of the world's most pressing demands regarding the availability of food, feed, fuel and materials.

Plant a radish.
Get a radish.
Never any doubt.
That's why I love vegetables;
You know what you're about!

Well maybe. These lyrics are from the 1960's play *The Fantasticks*. Performed continuously for 42 years, it was the world's longest running musical. The plot was simple, the songs charming.

But in the real world of plant breeding, life is far more complicated and 42 years is just a moment in time, as the researchers involved in crop genetic improvement at Pioneer Hi-Bred fully appreciate. Pioneer, a DuPont business, is the world's leading source of customized solutions for farmers, livestock producers and grain and oilseed processors. With headquarters in Des Moines, Iowa, Pioneer provides access to advanced plant genetics in nearly 70 countries. And hybrid seeds have never been in more demand.

Farmers around the world are trying to cope with climate changes and need plant varieties that can withstand droughts, floods, diseases and insects. And many farmers are shifting to crops tailored for biofuels—that is, fuels made from renewable resources like cellulose, starch or other plant matter. But conducting experiments to determine the performance of new hybrids takes a long time, requiring many years to study the potential of new hybrids in thousands of experiments conducted in different farm management conditions. For example, Figure 1 on the following page shows a side-by-side comparison of an old maize hybrid from the 1930s and a modern maize hybrid from the 1990s. The modern hybrids have a different plant structure that is more efficient at capturing key environmental resources (radiation, water and nutrients) and produces more harvestable yield than their older relatives.

In any given experiment, Pioneer researchers may have to wait days, weeks or months (depending on the objectives of the experiment) for plants to grow before they can examine the results. However, to accelerate the rate of progress and reduce the time it takes to create new hybrid seeds, one of the areas Pioneer researchers have turned to is "in silico modeling" methods—that is, modeling done using high performance computing (HPC) or in "virtual reality"—to simulate the predicted performance of new genetic combinations for different environments in advance of the experimental testing.

Competing for Acres

As environmental concerns and the price of petroleum make biofuels an attractive energy alternative, it may be more profitable for farmers to convert some of their farm land to the growing of plants for biofuels. At the same time, there are not a lot of new acres coming into production, and world demand for food is increasing, creating competition for use of the land. The solution, according to Pioneer, is to create a greater harvestable yield by growing more food on each acre of available

Figure 1 depicts an old maize hybrid from the 1930s (left) next to a modern maize hybrid from the 1990s (right).

land in a sustainable way. This creates "virtual acres"—as more crops are grown on each acre, a portion of the farmland can be reallocated for biofuel crops without lowering overall food production.

Mark Cooper, research fellow at Pioneer, explains the enormity of the challenge: "We can't do this with a single product tailored to a single environment—instead, we have to increase the genetic gain of the crops across a variety of different environments, especially as weather patterns continue their rapid shift. So we're trying to help farmers manage their risks by using our genetic knowledge to produce a broad spectrum of products that they can grow with assurance of a predictable yield and greater sustainability. The products have to have the right biochemical composition to meet the particular applications—whether it's for biofuels or human or animal consumption."

The Big Three—Genetics, Environment, Management

Pioneer researchers only know what a plant "is about" when they can understand the complexities associated with its genetics, as well as with the environment in which the plant will grow. They also must consider the methods by which farmers manage the land—when they plant, how much fertilizer is used, levels of irrigation, and so on. If one takes a more computational viewpoint, the company's crop breeding programs can be seen as search algorithms exploring the interaction between these three components. The new genetic product is a better search result found for that environment and management practice. These search strategies in modern commercial maize breeding programs involve multiple parallel searches conducted by multiple breeding programs with coordinated sharing of results to accelerate the rate of genetic improvement. HPC enables modeling of these applied search strategies to enhance the design of efficient breeding methods.

This is not an easy task—the amount of data that must be considered is truly astronomical. In particular, researchers are concerned with a plant's genotype—the set of genes carried by the plant—and its phenotype—the observable traits that the plant expresses as a result of the characteristics and traits coded in its genes (such as disease and insect resistant plants or sweet tasting corn). For example, anywhere from one to more than 200 genes control genetic variation for some of the traits of interest. For each of those genes, there can be two or more alleles (an allele is one member of the set of different forms of genes occupying a specific position on a specific chromosome). The number of total possible combinations of alleles for the corn genome is greater than the estimated number of all the atoms in the observable universe. Impossible to even imagine, let alone explore. Given that the experimental testing of all possible genetic combinations is not feasible, Pioneer wants to use modeling and HPC to leverage the genetics knowledge they have created from past experimental efforts and apply prediction methodologies to frontload future experiments with the most promising new genetic combinations. Pioneer's proof-of-concept work has indicated that over time, this will enhance the efficiency of experimental efforts and accelerate the rates of genetic improvement for the traits that contribute to successful corn hybrids.

"The volume of molecular data that is now available to us on the performance of genotypes and phenotypes has increased by orders of magnitude over the past 10 to 15 years," says Cooper. "Every genotype that Pioneer is testing has a genetic relationship to all the other genotypes we have worked with dating back to the 1920s—from a plant's great, great, great grandparents to today's progeny. These family trees are deep, and the associated trait genetics knowledge is rich in the information that the plant breeders require to make further genetic improvements to corn hybrids. There is invaluable information contained in these data sets, but it's useless if we can't organize and analyze it. But even if we have the data sets we need to solve a specific problem in the life sciences, the time it takes to analyze all that information has also increased by orders of magnitude."

Finding the Sweet Spot

Starting with the farmer's environment and management practice, their task is to find the selection trajectory that gives rise to the best combination of genes to allow those genetic characteristics to achieve their full potential. They are, as Lane Arthur, DuPont business manager for computing technology, put it, "looking for the sweet spot."

Take hybrid corn, for example. Pioneer performs exhaustive testing of a large number of genetic combinations in thousands of experiments each year. When they decide to commercialize a new corn hybrid, they want to know if the behavior they have seen in their experiments is representative of the environmental and land management conditions the product will encounter in the future. If the corn is going to be used for biofuels, the combination of the big three—genetics, environment and management—should yield corn that has a higher extractable starch content, the ingredient used to make ethanol.

The company's breeding program ideally produces a new maize or corn hybrid with improved yield and agronomic performance for use by many farmers across thousands of farming acres. Also, the product sold to farms goes beyond bags of seeds; Pioneer provides them with the knowledge they need to be successful—knowledge the company has gained through decades of experimentation.

Handling Bushels of Breeding Data with HPC

Managing this huge influx of information is absolutely essential for Cooper and his fellow researchers to move the state of the art in plant breeding along. They need a way to digest, analyze and turn around the results of these analyses rapidly and efficiently. Finally, the Pioneer researchers have to perform the difficult act of predicting the future in their field of applied genetics. Mathematically, the number of gene combinations they need to investigate makes it difficult just to manage these data, not to mention performing valid statistical analyses. HPC is critical to getting the job done.

Pioneer has a long history of working with powerful computer systems, starting with early mainframes, then moving on to groups of minicomputers, and finally, in 2000, the implementation of HPC clusters using off-the-shelf components. These clusters have been continuously upgraded to keep pace with the ongoing deluge of data and complex simulations required to explore the characteristics of various hybrids as they evolve over time.

"The fact is that we do not have the resources, the time and the land needed to experimentally explore all the possible avenues of research we would like to cover," Cooper explains. "This is where HPC is invaluable. The power of our HPC clusters allows us to simulate a wide variety of parameters relating to plant breeding at the cellular, plant and environmental levels. It's given our researchers the tools they need to get answers immediately rather than waiting several weeks for their data to be processed. These answers drive our plant breeding strategy and have a direct impact on our competitive position in the marketplace."

Pioneer uses its HPC capabilities in two primary ways. On one hand, the clusters serve as a high-powered batch processing system that crunches huge amounts of data and then applies statistical analyses to these data. On the other hand, the same HPC cluster configuration is used for modeling and simulation of the breeding process. This allows Pioneer to determine which direction is likely to bring the most positive results in terms of genetic improvement. Software visualization capabilities allow the researchers to work more easily with massive amounts of data.

The results of both approaches—number crunching and modeling/simulation—have allowed the researchers to engage in deep theoretical work to study the genetic architecture of many traits of corn that is now being put into practice. As Cooper says, "We're essentially starting to write the next textbook of applied genetics in plant breeding."

HPC Competitive Impact: Food, Feed, Fuel and Materials for the Future

The supercomputing capabilities of Pioneer have allowed the company to boost the quality of potential new seed hybrids, some of which are on the path to commercialization.

Their HPC clusters not only allow the researchers to analyze massive amounts of data on plant genotypes and phenotypes, they also provide answers an order of magnitude faster than the company's previous computer systems. Also, because the clusters can work with astronomical numbers of gene combinations and weave in environmental and management data, the researchers now have a modeling platform from which to predict the future—they can make informed estimates of the outcomes of specific breeding programs and move ahead based on these data.

Cooper adds, "HPC has become such an enabling resource that it has stimulated innovation among our researchers. These are very smart people, and now they are getting answers back in hours instead of days and weeks—in many cases it has totally changed the way they think about the problems they are working on. On the modeling and simulation side, I know Pioneer is way out in front of the curve, and our scientists are taking advantage of the opportunities that these enabling technologies provide."

HPC is giving the company the answers it needs to shorten the time to market for new products, stay out in front of its competitors, and provide the agricultural products that have the potential to help alleviate world food problems and provide alternative sources of energy. Their research can help create a greater harvestable yield on existing acres and create plants that can be grown as components of sustainable farming systems when more inhospitable environments are encountered.

"These answers drive our plant breeding strategy and have a direct impact on our competitive position in the marketplace."

Mark Cooper, Research Fellow, Pioneer Hi-Bred

"HPC lets us run a breeding program that is 10 to 50 times the size of what would be possible otherwise," says William S. Niebur, DuPont vice president, Crop Genetics Research and Development. "Our resource needs are threefold—software, hardware and grey ware, or brain power. When these critical resources are combined and focused on the genetic improvement of crops, the end result is new plant products that are better able to cope with water and nutrient deficiencies, insects, diseases and competition from weeds."

In Brief

Key Challenges

- Manage and analyze huge amounts of molecular data on plant genotypes and phenotypes to chart a future course of action for Pioneer products
- Help farmers maintain a balance between growing plants as an alternative energy source while continuing to produce the food, feed, fuel and materials that the world needs
- Provide robust products for farmers and grain processors that increase their profitability and provide overall sustainability for our environment

Solutions

- Use high performance computer clusters for large-scale number crunching and statistical analysis of the data, and modeling and simulation to determine the outcome of genetic, environmental and land management factors on potential new seed products

Key HPC Benefits

- Manages, analyzes and extrapolates massive amounts of genetic and environmental data for prediction of new hybrid seed development
- Improves quality of current hybrid seeds now under testing for possible commercialization
- Speeds up answers to research problems by an order of magnitude—from days and weeks to a matter of hours
- "Predicts the future" of seed breeding programs
- Boosts quality and quantity of potential new seed products
- Shortens time-to-market for new products
- Stimulates innovation and productivity among the research staff
- Helps alleviate world food and fuel problems by increasing the harvestable yield of crops on existing acreage

Web Site

- www.pioneer.com

Council on Competitiveness
1500 K Street NW, Suite 850, Washington, D.C. 20005 T 202 682 4292
Compete.org