CASE STUDY
Intel® Xeon® Processor E5-4650L Product Family
High-Performance Computing
Big Data Analytics

Look Inside.™

Understanding life on earth

TGAC uses high-performance computing solutions from SGI powered by the Intel® Xeon® processor E5-4650L product family to sequence and assemble one of the most complex genomes, the bread wheat genome

The bioinformatics team at The Genome Analysis Centre (TGAC) in the U.K. uses the SGI UV2000* high-performance computing (HPC) system, powered by the Intel® Xeon® processor E5-4650L product family, in its work to promote a sustainable bioeconomy. Using the largest SGI UV installation in the U.K., researchers categorize, process, and analyze genome sequences of various crops, animals, and microbes, including bread wheat, which has a genome sequence five times more complex than humans. Its work is essential in securing future food supply.

Challenges

- **Sustainable bioeconomy.** One of TGAC's primary goals is to understand crop genomes so new varieties can be developed to secure food supply in the face of a growing population and environmental change.
- **Computational power.** TGAC's research output requires significant computational effort to categorize, process, and analyze genome sequences of various plants, animals, and microbes.
- **Big data.** DNA analysis produces immense data sets that have increased nearly 200-fold within six years and which continue to grow.

Solution

- **SGI and Intel.** TGAC has deployed the largest SGI UV2000 HPC system in the U.K. (and the third-largest globally) running a Linux* operating system, powered by the Intel Xeon processor E5-4650L product family.

Technology results

- **Memory capacity.** The SGI UV2000 installation at TGAC has 2,560 cores and 20 TB of coherent main memory for in-memory computing in a single-image system, and can scale to 4,096 cores and 64 TB of RAM.
- **Workflow consolidation.** The Intel® technology-powered SGI UV2000 enables TGAC to consolidate complete workflows on a single system, and offers a very low IT burden per compute core versus comparable clusters or scale-out systems.

Business value

- **Research capability.** TGAC's scientists can assemble large, complex genomes that would be extremely difficult to achieve on smaller systems.
- **Wheat genome.** TGAC's scientists sequenced 17 of the 21 chromosomes in the bread wheat genome and have generated all of the related chromosome assemblies.
- **Improved crops.** Once the full sequence is available, scientists will be able to examine how genes control complex traits such as yield, grain quality, disease, and pest resistance.

TGAC The Genome Analysis Centre™

"Researchers estimate that the full wheat genome sequence will be available within three years, which will be a vital resource for improving crops. Without the computational power we have from SGI and Intel, it would be unlikely to happen in our lifetime."

Timothy Stitt,
Head of Scientific Computing,
The Genome Analysis Centre

A scientific center of excellence

Researchers and computer scientists at TGAC apply state-of-the-art genomics and bioinformatics to research, analyze and interpret multiple, complex data sets. TGAC receives strategic funding from the U.K.'s Biotechnology and Biological Sciences Research Council (BBSRC), and is a member of the Norwich BioScience Institutes (NBI), situated on the Norwich Research Park.

TGAC also hosts one of the largest HPC facilities dedicated to life science research in Europe. Timothy Stitt, head of scientific computing at TGAC, explains the central role that technology plays in TGAC's work: "All our research output requires significant computational effort to categorize, process, and analyze genome sequences of various organisms, many of which are incredibly complex. Not only do we have intense computational needs, our storage requirements are enormous."

His colleague Paul Fretter, the science computing team leader at NBI, agrees. "There has been a revolution in data-generation technologies, which have transformed the landscape of life sciences. The technology used to extract DNA data is changing rapidly and the rate of information output is accelerating fast. Between 2008 and 2014, we saw a nearly1,000-fold increase in information yield from laboratory sequencing instruments. HPC is fundamental to our work—without it, there would be no genome analysis as we know it today."

Big data every day

One of the main research priorities at TGAC is to apply genomic technologies to maintaining food security in the face of a growing population and environmental change. As part of this role, TGAC is a key participant in the International Wheat Genome Sequencing Consortium (IWGSC) project, which generates and analyzes sequences from the bread wheat genome.

The wheat genome project is one of the more computationally intensive scientific and research challenges that TGAC has undertaken. The genome for bread wheat is particularly large

Bringing together big data analytics, HPC, and world-class genomics research

and complex; it is at least five times larger than the human genome and contains many repeated sequences. Bread wheat also has three distinct ancestral sub-genomes, so trying to sequence and assemble the bread wheat genome is as difficult as sequencing and trying to interpret the genomes of a human, a chimpanzee, and a gorilla all at the same time.

Like all aspects of TGAC's work, the wheat genome project is a perfect example of big data analytics in action. Laboratory data comes from high-throughput sequencers that analyze the physical matter DNA. After this primary analysis, the data is interpreted to read out the sequence of letters that represent each strand of DNA. It is then submitted to quality control, assembly, and annotation, where TGAC's scientists can then start to interpret this data in order to understand the role of each part of the genome.

Richard Leggett, project leader for quality control and primary analysis at TGAC, explains the assembly process: "We think of a 'genome' as a string of millions or billions of letters that represent four basic biological compounds—the wheat genome, for example, is represented by a string of 17 billion characters. But the most common DNA sequencing machines can only 'read' around 100 to 300 letters of DNA at a time, so when we sequence a genome we have to split it up into lots of smaller chunks. Assembly is the process of putting them back together again; unfortunately, there is no way to know where in the genome each sequenced chunk comes from. It's a bit like taking 30 copies of a novel, cutting up all the words, putting them together in a big pile and then trying to re-create the novel. It requires a lot of computing power."

The bioinformatics team at TGAC also conducts data-intensive re-sequencing projects that compare and contrast the genomes of different individuals within a population of the same species of plant or animal. They also perform expression analyses to understand how organisms respond to changes in their environment and metagenomics analyses, which looks at the DNA of mixed species—for example, from a sample of soil containing all kinds of microscopic bacteria or fungi. "Metagenomics is like trying to understand what's left over after hundreds of different novels have been shredded and dumped in a pile," says Leggett.

Lab operations can process between 2 and 4 TB of data per week—approximately 2 TB of which then needs to be stored. One of the issues that intensifies the need for high computational power is that much of the information is being analyzed for the first time, which is known as de novo sequencing.

As Stitt points out, "When the team runs an analysis for the first time, it is sometimes not obvious which algorithm or parameter set will produce the best result without first trying out a number of different combinations."

Memory and performance needs

The work at TGAC has led to the development of a number of new applications and methods for assembling and analyzing large sequencing data sets, the majority of which are written to run on a Linux operating system. The right hardware is crucial, and TGAC has taken delivery of its third SGI UV system powered by the Intel Xeon processor E5-4650L product family.

Fretter was involved in the decision to work with SGI and explains the reasoning: "We first heard about SGI UV through conversations with our Intel representative. SGI has a great reputation for making HPC hardware work at optimal levels and it soon became clear that the UV had the features, notably the memory, necessary for our work."

The UV2000 is a large shared-memory machine, which, in combination with the Intel Xeon processor E5-4650L product family, makes available up to 4,096 cores and 64 TB of coherent main memory for in-memory computing in a single-image system.

Stitt adds, "With the SGI UV2000, we have the largest installation of its kind in the U.K., and we can consolidate complete workflows in a single system. It has a very low IT burden per compute core, in contrast to comparable clusters or scale-out systems. It also provides 20 TB of RAM that greatly increases the size of problems we can handle at one time, while simplifying the programming required to do it."

He continues, "The result is that our scientists can assemble large genomes, such as wheat, that would be extremely difficult to achieve on smaller systems."

Feeding tomorrow's population

TGAC's scientists have now sequenced and assembled 17 of the 21 chromosomes of the wheat genome, using the Intel technology-powered SGI UV2000 system for the sequencing work.

Stitt points out, "Our work with the IWGSC is just one of the many success stories coming out of TGAC. Researchers estimate that the full wheat genome sequence will be available within three years, which will be a vital resource for improving crops. But without the computational power we have from SGI and Intel, it would be unlikely to happen in our lifetime. The simple fact is that we could not do the work we do without the computational facilities we have. When it comes to HPC in the Biosciences, this is as good as it gets."

Once the full sequence is available, scientists will be able to identify how genes control complex traits such as yield, grain quality, disease, and pest resistance. By studying wheat's internal structure, they will also gain insights into how traits important for pest resistance and tolerance of drought and other environmental stresses have developed. Plant breeders can use that information to produce new wheat varieties with higher yields and improved sustainability that are better able to feed the global population.

Christine Fosker, head of research faculty office at TGAC, explains further, "Conventional crop breeding has been used for generations to always attempt to create the most desirable forms of crops through cross-breeding. Because our work gives us unparalleled insight into how and why organisms work at a fundamental level, we can give breeders more information and therefore much more precise tools that will take a lot of the guesswork out of time-consuming crop development. By using sequencing to survey and identify emerging pathogen threats we could, for example, also give farmers rapid information on which crops are most resistant to local pathogens that could otherwise go on to cause disastrous crop failures. It all has huge implications on our ability to feed ourselves as a planet."

TGAC's work is not confined to the wheat genome project. It is currently also looking at minimizing the impact of ash dieback, a chronic fungal disease affecting ash trees in the U.K., by analyzing the genomes of the relevant trees as well as the fungi causing the disease. Other areas of the organization are looking at what changes could be made to overcome growing levels of antibiotic resistance in the population.

Fosker says, "Having the genome sequence is just the start. Once that's in place, we can start effectively asking questions of it, by looking at differences and trends, tracking changes over time, and pinpointing the smallest anomalies that could potentially tell us so much. That's when we really start to fully understand life on earth in all its diversity."

Spotlight on TGAC

TGAC is a member of a research campus partnership that also includes the John Innes Centre, the Institute of Food Research, the Sainsbury Laboratory, the University of East Anglia, and Norfolk and Norwich University Hospital. With its world-class expertise in data analytics, genomics, geochemical cycles, crop biology, health, and nutrition, the partnership addresses global challenges of food and energy security, sustainability and environmental change, and healthy aging. TGAC is strategically funded by BBSRC and operates a National Capability to promote the application of genomics and bioinformatics to advance bioscience research and innovation.

0814/JNW/RLC/XX/PDF 331068-001EN