TECHNOLOGY BRIEF

Cray helps developers of the Ray De Novo genomic assembler software tune their application — and make gains on assembling a human genome in a single hour

Organizations

Université Laval
Quebec City, Quebec, Canada
www.ulaval.ca

Cray Inc.
Seattle, WA
www.cray.com

Cray XE6™ Supercomputer

The Cray XE6 supercomputer features two innovative technologies: AMD Opteron™ 6300 Series multicore processors and the custom-designed high bandwidth, low latency Gemini interconnect. The result is a system that brings production petascale to a wider high performance computing community and fundamentally changes how Cray systems communicate.

Each Cray XE6 blade includes four compute nodes for high scalability in a small footprint. Each compute node has two AMD Opteron 16-core processors coupled with its own memory and Gemini communication interface and is designed to efficiently run up to 24 MPI tasks. Alternately, it can be programmed to run OpenMP within a compute node and MPI between nodes. Each Cray XE6 node can be configured with 32 GB or 64 GB DDR3 memory.

The Ray Development Team

The Ray development team is led by Jacques Corbeil, a full professor in the molecular medicine department at the Faculty of Medicine at Université Laval. Dr. Corbeil leads research projects at the Centre de Génomique de Québec aimed at understanding the host-pathogen interactions, more specifically those involved in HIV infection, legionellosis, viral respiratory infections and leishmaniosis, using genomic tools and bioinformatics.

Cray Inc.
901 Fifth Avenue, Suite 1000
Seattle, WA 98164
Tel: 206.701.2000
Fax: 206.701.2500
www.cray.com

The Project

Ray is a highly parallel computer software developed at the Université Laval that performs de novo genome assemblies with next-generation DNA sequencing data. Written in C++, Ray can run in parallel on numerous interconnected computers using the message-passing interface (MPI) standard.

The goal of the Ray project is two-fold: to provide a foundation for distributed software in genomics; and continually reduce the time for genome assembly. Ultimately, science intends to assemble a human genome in less than one hour. An essential component to realizing these goals is leveraging the architecture underneath the software. To that end, Université Laval researchers collaborated with Cray, using Cray hardware and applications expertise to optimize their code.

The Reason

Next-generation sequencing (NGS) — a recent advancement in DNA sequencing technology — has resulted in a new class of devices that allow for the analysis of genetic material with unprecedented speed and efficiency. However, the step in the workflow that generates raw genomic data is outpacing the rate the data can be analyzed.

NGS produces large quantities of small fragments of DNA called "reads" that must be assembled into a useful form. To put into context, the standard benchmark human set is 6 billion pair reads. In general, transforming these small reads is done by either assembling individual reads (de novo) or mapping these pieces against a reference genome (mapping). Applications that perform de novo assembly are well-suited for high performance computing (HPC) systems, in particular. As NGS-generated data continues to grow exponentially, researchers need tools that can accurately and efficiently make use it. These factors make optimizing de novo assemblers like Ray extremely important.

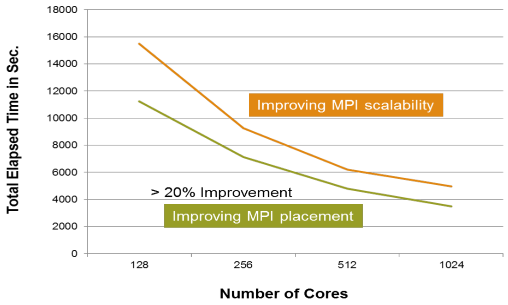
The Results

Efficient scalability is critical in order to reduce the total time for assembly and Ray is implemented using message passing interface to leverage distributed-memory architectures. The Cray XE6 system provides the necessary scalable architecture built on the custom-designed high bandwidth, low latency Gemini interconnect.

Using a human gut microbiome benchmark, the team investigated the performance characteristics of Ray using a Cray XE6 system with 16-core AMD Opteron processors and 32GB of memory per node to determine areas for optimization. The findings showed biodirectional extension of seeds was the most time-consuming step. Optimized MPI task allocation resulted in a performance improvement of approximately 20 percent. Overall, the Ray assembler scaled well, showing scalability to beyond 1,024 cores on the Cray XE6 machine. The favorable results have prompted the collaborators to conduct further investigations, including an examination of Cray's next-generation Cray XC30™ architecture.

Human gut gene catalog
Metagenomics
124 Individuals, 577 GB generated
Beijing Genomic Institute

Application Tuning


Total Elapsed Time in Sec.
Number of Cores
Improving MPI scalability
> 20% Improvement
Improving MPI placement