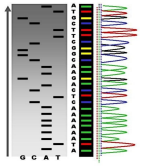
High Performance Genome Assembly on a Cray Supercomputer

Mikhail Kandel[1,2], Steve Behling[1], Bill Long[1] and Carlos P. Sosa[1,3], [1]*Cray Inc*, [2]*University of Illinois and* [3]*BICB, University of Minnesota Rochester*
Sébastien Boisvert and Jacques Corbeil, *Département de Médecine Moléculaire, Université Laval*
Lorenzo Pesce, *Computation Institute, University of Chicago*

Project Objective

- Assemble a human genome in approx. 1 hr. The standard benchmark human set is 6 billion pair reads
  - Currently it takes approximately 11 h with Ray and approx. 512 cores
  - It takes much longer with other assemblers, e.g., 3-6 days. Many are not well parallelized or even parallelized

Introduction



An example of the results of automated chain-termination DNA sequencing
Courtesy of: http://en.wikipedia.org/wiki/DNA_sequencing.

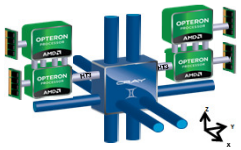
- Recent progress in DNA sequencing technology has yielded a new class of devices that allow for the analysis of genetic material with unprecedented speed and efficiency
- This new tech is referred as Next Generation Sequencers (NGS)
- By breaking up DNA into millions of small strands (20 to 1000 bases) and reading them in parallel, the rate at which genetic material can be acquired has increase by several orders of magnitude
- The technology to generate raw genomic data is becoming increasingly fast and inexpensive when compared to the rate that this data can be analyzed
- Assembling small reads into a useful form is done by either assembling individual reads (*De Novo*) or mapping these pieces against a reference (mapping)
- Ultimately the ability to assemble a human genome in the order of an hour would be ground breaking
- In this work we present our experience with applications such as Picard, GATK, BWA, SAMTOOLS, BLAST and others for whole genome and exome sequencing
- We specifically highlight our work on HPC systems, particularly with Ray, a parallel short-read assembler code

Methodology

- Efficient scalability is critical in order to reduce the total time for assembly
- Using a Human Gut microbiome benchmark, the performance characteristics of Ray were investigated on a Cray XE6 supercomputer to determine areas for optimization
- The Cray system provides a scalable architecture built on a custom designed high bandwidth, low latency Gemini interconnect in a 3D torus topology

Cray XE6 Hardware
2.1 GHz AMD Interlagos
32 GB of Memory/node

Compute Node with Gemini Interconnect



Ray Hybrid *De Novo* Assembler

- The goal in *De Novo* assembly is to correctly assemble short reads into longer sequences
- Current Next Generation Sequencing technologies offer increase in throughput and decrease in cost and time
- Ray has been developed to assemble reads obtained from a combination of sequencing platforms
- The algorithm implemented in Ray is based on de Bruijn Graphs
- It is implemented using MPI to leverage distributed-memory architectures such as the CRAY XE6
- The fastest current time to assemble a human genome of 6 billion pair reads is approximately 11 hours using Ray on 512 cores.
- Other assemblers require several days to complete due to limited parallelization and scalability

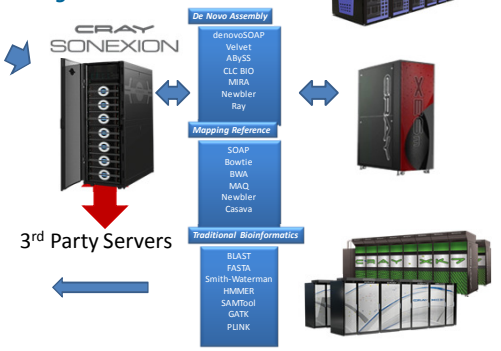
S. Boisvert, F. Laviolette, and J. Corbeil, J. Comp. Biol. 17, 1519-1533(2010)v
S. Boisvert, F. Raymond, E. Godzaridis, F. Laviolette, and J. Corbeil, Genome Biology 2012, 13:R122

Acknowledgements:
- Many thanks to Per Nyberg for valuable suggestions
- Biomedical Informatics and Computational Biology Program, UMR, Rochester
- Cray Inc
- Compute Canada

S.B. is recipient of a doctoral award from the Canadian Institutes of Health Research (200910GSD-226209-172830). JC is the holder of the Canada Research Chair in Medical Genomics

Next Generation Sequencing Tools

Cray Storage and Data Management Solutions



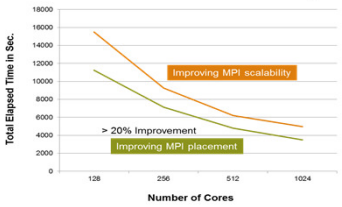
Cray Computing Solutions

Data Analysis

Results

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

Application Tuning



- Bidirectional extension of seeds is the most time consuming step
- Optimized MPI task allocation resulted in a performance improvement of approximately 20%
- Ray shows good scalability beyond 1024 cores

Summary and Future Work

- *Ray De Novo* assembler is a well designed and scalable application
- The goal of optimization work is to continuously reduce the time for genome assembly
- As a first step of our investigation, Ray has scaled to beyond 1024 cores on the Cray XE6
- Further investigation for optimization opportunities will be pursued including an examination of the benefits of Cray's next generation XC30 architecture
- A tightly integrated NGS solution that considers the end-to-end workflow is also being investigated