Cray Demonstrates Top-Level Performance and Scalability on Very Large Datasets with Velvet

Velvet: *de novo* assembly using very short reads

Velvet is a de novo genomic assembler specially designed for short read sequencing technologies such as Solexa or 454. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI) in the U.K., Velvet currently takes in short read sequences, removes errors, and produces high-quality unique contigs. It then uses paired-end read and long-read information, when available, to retrieve the repeated areas between contigs.

Cray® CS300™ LMS Supercomputer

The Cray CS300 large memory system (LMS) delivers direct access to many terabytes (TB) of memory for workloads that do not have the need for a large number of processors. The CS300 LMS is designed for extremely high RAM requirements that can't be addressed by traditional multi-socket servers.

Preconfigured as a ready-to-go solution, the CS300 LMS is based on optimized configurations utilizing the same building block platforms as the Cray CS300-AC system and incorporating vSMP Foundation™ software from ScaleMP™. This solution offers users a single point of management and application scalability for managing extremely large amounts of structured or unstructured data in-memory with fast turnaround for large data research.

The Problem

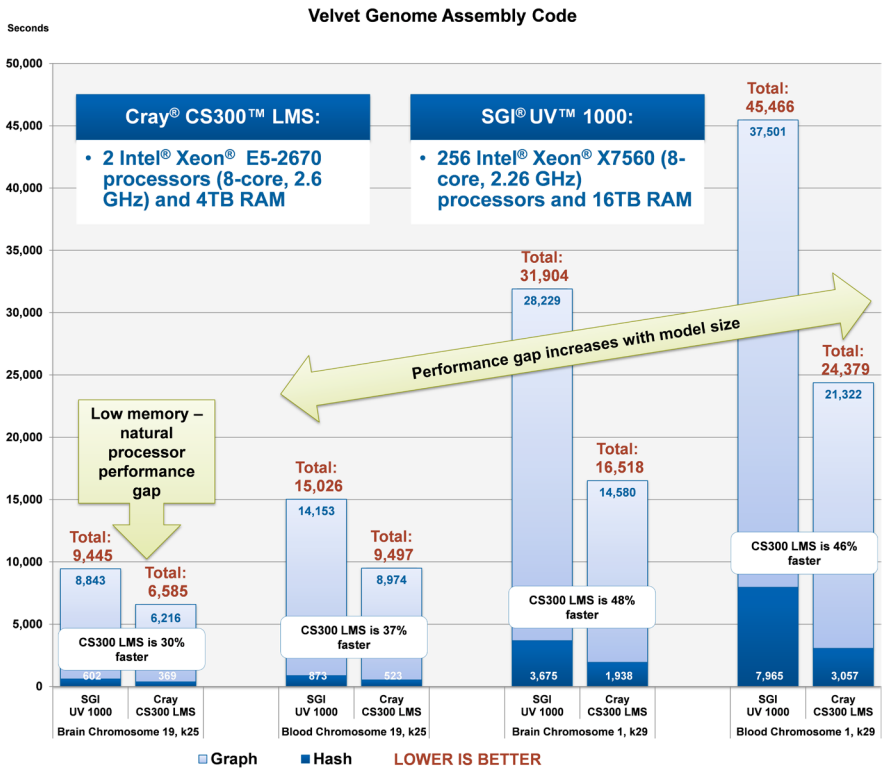
Next-generation sequencing (NGS) describes the modern DNA sequencing technologies that allow for the analysis of genetic material with unprecedented speed and efficiency. Its advent is shifting genomic and molecular biology research from a problem of laboratory-based chemistry to one well suited to high performance computing (HPC).

In simple terms, NGS involves breaking up DNA into millions of small strands (20 to 1,000 bases) and then reading them with a computer. The rate at which genetic material can be acquired has increased by several orders of magnitude. In general, assembling small reads into a useful form is done by either assembling individual reads (de novo) or mapping these pieces against a reference (mapping).

Velvet is a de novo genomic assembler designed for short reads generated by NGS sequencers. One of the computing challenges of running the application is its high demand for very large amounts of memory accessible from an OpenMP application programming interface (API).

The Solution

Tuned for applications requiring a low core count but very large amounts of shared memory, the CS300 LMS system delivers up to 8.75 terabytes of memory. The system incorporates vSMP Foundation™ software from ScaleMP™ and is designed to reduce overall system cost and maximize performance and memory capability. A single CS300 LMS system allows applications to access all of the memory (physically in multiple systems) without the complexity of a cluster and allows large shared memory applications to run at maximum performance and scale linearly, based on the amount of memory.



Velvet results on a CS300 LMS: Velvet ran with two different chromosomes and two different k-mer sizes. As the size of the problem increases and the amount of memory per node is exceeded, Velvet can exploit the power of fast access to high speed memory via vSMP Foundation. Additionally, as the size of the system increases so does the gap in performance when compared to a traditional shared-memory system.

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