Genome Assembly and Parallel Computers

University of California, Berkeley
Ouliana Panova

I. Short Biography

I am a first year PhD student in Materials Science and Engineering. I did my Bachelor's and my Master's at the University of Arizona, with an emphasis on computational materials for the latter, and my thesis dealt with automating the analysis of dendritic metal micrographs using computer vision techniques. My current research, although still in its infancy, will involve the modeling of liquid structures of metals and their effects on their thermodynamic properties in solar thermal applications.
I have enrolled in this class because I believe that for any modern applications all code should be as parallel as possible; and besides, it is a very useful skill to have in general.

II. Parallel Problem

The problem chosen here is the parallelization of a very popular one: genome sequencing. While the basic concept of reading the linear sequences of the four DNA nucleotides A, T, C and G that make up the genes and is rather straightforward, it is a non-trivial matter to implement such a mapping.
The main method for genetic sequencing is called Sager sequencing and involves breaking the long strands of DNA into small pieces that can be read individually [1]. The quanta of information carried by the genes is a *base pair*, which contains two complimentary nucleotides; each of the strands that are analyzed have on average 700 to 800 of such pairs, while the full genome of a human has over three billion. This method, although fast and effective in the reading of the nucleotide sequence, has a main drawback – the DNA being split is done so at random and thus the strands have to be somehow reassembled afterwards in order to obtain any useful information. Such assembly relies on finding overlaps between the fragments and thus linearly assembling them together; additionally, in order to account for as much information as possible, the DNA is first cloned multiple times before breakdown which, while enhances the data, greatly increases the amount of fragments to analyze and align. The method proposed by Kalyanaraman et. al. [2] is an effective parallelization of the assembly of the genome from the fragments obtained throught the Sanger technique.
While the DNA is profusely cloned, there still remain many areas that have not been sequenced; such areas divide the genome into many *contigs*; this division is favorable to parallel computing as it naturally divides the problem into several independent batches that can be analyzed by different processors. Thus, clusters are intellegently generated by a comparison of matches between the fragments; each cluster is then individually assembled into a contig and stored. This process is illustrated in Fig. 1.
The analysis of the clusters is, however, not the only opportunity for parallelism; the division

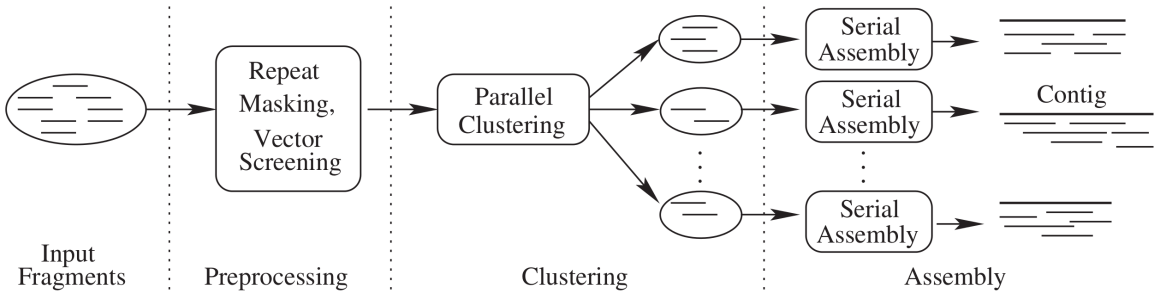

Figure 1: Illustration of the clustering algorithm [2].

into clusters can itself be parallelized as well. The algorithm starts with only one fragment per cluster, determines which clusters are "promising pairs" and tries to align them; if the alignment is successful, the two clusters are merged together, otherwise they remain separated and the process is repeated until there are no more promising pairs to try. The problem is taken on by a master processor and multiple workers; the master distributes pairs to be aligned and collects pairs of promising matches provided by the workers; in turn, the workers provide the pairs as well as the results of the alignments. This hierarchy is illustrated in Fig. 2.

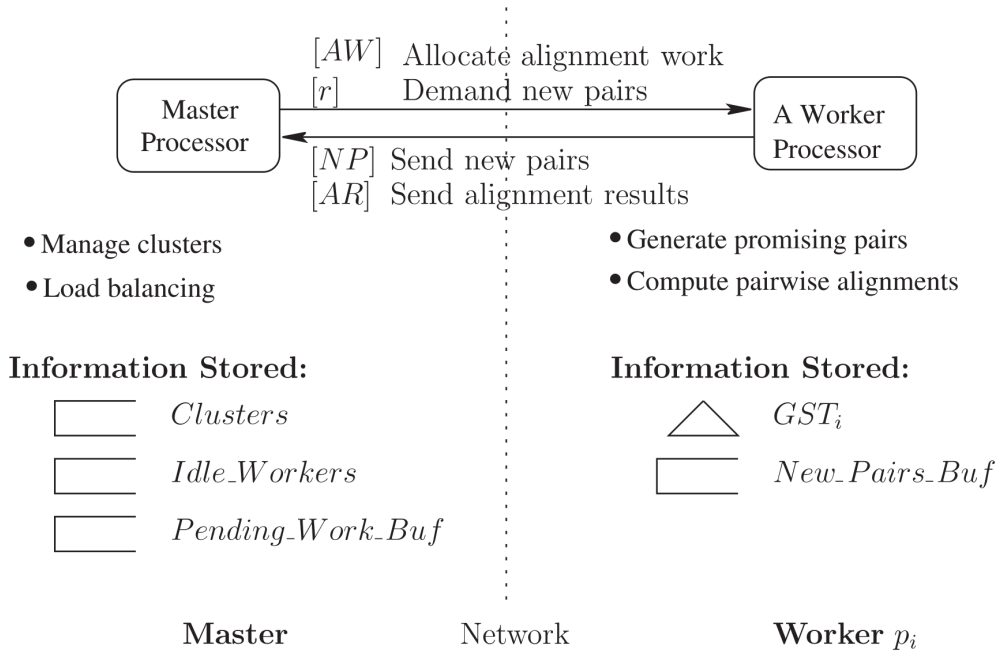

Figure 2: The Single-Master-Multiple-Workers hierarchy [2];

The implemetation of this parallel algorithm on the 1024 BlueGene/L supercomputer (ranked 362 on the Top500 as of November 2012) reduced sequencing times from several weeks to less than 24 hours [2], and completely automated the sequencing process. It has been implemented on genomes of important crops such as maize and has been proven to show very little error in the provided results.

References

[1] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[2] A. Kalyanaraman, S.J. Emrich, P.S. Schnable, and S. Aluru. Assembling genomes on large-scale parallel computers. *Journal of Parallel and Distributed Computing*, 67(12):1240–1255, 2007.