Next Generation Bioinformatics Tools on Next Generation Hybrid Architectures

Eduardo Ponce
National Institute for Computational Sciences
1 Bethel Valley Road
Oak Ridge, TN 37831-6173
eponcemo@utk.edu

Bhanu Rekepalli
National Institute for Computational Sciences
1 Bethel Valley Road
Oak Ridge, TN 37831-6173
brekapal@utk.edu

Gregory D. Peterson
National Institute for Computational Sciences
1 Bethel Valley Road
Oak Ridge, TN 37831-6173
gdp@utk.edu

ABSTRACT

The life sciences are experiencing the need of novel analysis tools to handle the massive flow of information continuously added to our biology knowledge base, due to next generation sequencing technologies. One such case is genomic sequence searches for characterizing an organism's genetic makeup which provides invaluable insights to associate phenotypic changes with functional and evolutionary relationships; having high-impact on health and medicine since it allows understanding response activity from proteins and functional RNA when in presence of particular diseases and drugs. Consequently, controlled genetic manipulation offers disease-resistant organisms, effective drug treatments, and other biomedical enhancements. This research describes the Highly Scalable Parallel Hybrid (HSPH) bioinformatics applications for fast genomic sequence analysis using high-performance hybrid architectures composed of Intel Xeon Phi coprocessors. HSPH's framework, adapted for symmetric execution, is presented as a case study of parallel software wrappers which uses resourceful workload distribution, I/O management, fault tolerance, and checkpoint recovery techniques suitable for bioinformatics applications. Weak and strong scaling experiments demonstrate linear scalability to thousands of processors. We will be discussing implementation and scaling studies run on two HPC architectures; one is Beacon at the National Institute for Computational Sciences and the other Stampede at the Texas Advanced Computing Center. Beacon is an Intel Many Integrated Core cluster containing 48 compute nodes each with two 8-core Xeon and four 60-core Xeon Phi processors targeted to understand the performance and energy impact of next generation processors and memories on scientific implementations. On the other hand, the Stampede supercomputer has 6,400 compute nodes equipped with two 8-core Xeon and one 61-core Xeon Phi processors, and NVIDIA Kepler GPUs with the goal of advancing science and engineering research that involves computational and data-intensive operations. Along with discussing how we could adapt our HSP approach to other informatics applications, we present the idea of integrating these applications into automated workflow systems which provides immediate access to a unique combination of powerful parallel bioinformatics analysis tools on supercomputing platforms for large-scale knowledge discovery.

Keywords

Bioinformatics; sequence alignment, high-performance computing, parallel processing.