Abstract

Over the years, paradigm shifts have occurred in bioinformatics research because of advanced sensor technology such as next generation DNA sequencer and X-ray diffraction. For example, new massively parallel DNA sequencing platforms have reduced the cost of DNA sequencing by over two orders of magnitude and have the potential to dramatically accelerate biological and biomedical research. However, the volume of data produced by such scientific sensors is immense and bioinformatics researchers are demanding new methods for handling these massive amounts of data. This project aims at establishing new computing platform with what we call ‘Stream Computing’. Stream Computing enables us to analyze incoming data from sensors with ‘on-the-fly’ model rather than ‘store-and-processing’ model. This real-time processing will enable highly efficient usage of network and storage model, and also the bioinformatics researchers will obtain their required results at a time when their scientific experiments are finished.

Stream Computing and Bioinformatics

IBM Research has developed a Stream Computing System called System S. Stream Computing is a new computing paradigm that enables faster-than-real-time data analysis by sequentially processing massive amounts of incoming data in memory, extracting the significant analytic results without storing the data. This contrasts with the existing batch-type processing platforms that deal with accumulated data. In this project, we will build a new computing infrastructure on top of System S that will handle the massive amounts of data produced by DNA sequencers and other biological sensors, and thus provide bioinformatics researchers with a new analytic infrastructure. This infrastructure may support another paradigm shift with the potential of contributing to biological and biomedical research that better exploits next-generation experimental devices such as these DNA sequencers.

Stream Computing System (IBM System S)

StreamDNA PROJECT
– Stream Computing for Next Generation DNA Sequencer –

Hiroya Matsuura¹ and Toyotaro Suzumura²
Tokyo Institute of Technology
IBM Research – Tokyo

References
Lisa Amini, ‘SPC: A Distributed, Scalable Platform for Data Mining’

Contacts
Toyotaro Suzumura : suzumura@es.titech.ac.jp
Hiroya Matsuura : matsuura.ha@tm.titech.ac.jp
Suzumura Laboratory : http://sites.google.com/site/tokyotechsuzumuralab/ (Japanese Only)