A cloud framework for high troughput biological data processing

The molecular systems biology community has to deal with an increasingly growing amount of data. A recent programming model that addresses the data deluge is MapReduce which enables easy processing of huge data volumes on large sets of computing resources. However, the availability of appropriate local computing resources is often limited. Cloud computing addresses this issue by providing virtually infinite resources on demand, usually following a pay per use model. In this paper we present our cloud based high throughput computing infrastructure (VCE) which combines the Software as a Service (SaaS) approach with the MapReduce programming model for data-intensive applications, and a configurable distributed file system as provided by the Hadoop framework. Within this infrastructure we realized an application in the field of molecular systems biology which maps tryptic peptide fragmentation mass spectra data against a large scale mass spectral reference database (Promex). We evaluate this application on a local cloud resource and study the effects of different configuration parameters as provided by the application, the Hadoop framework, and the available computational and storage resources.

Primary authors : Mr. KOEHLER, Martin (University of Vienna) ; Mr. KANIOVSKYI, Yuriy (University of Vienna)

Co-authors : Prof. BENKNER, Siegfried (University of Vienna) ; Dr. EGELHOFER, Volker (University of Vienna) ; Prof. WECKWERTH, Wolfram (University of Vienna)