



LAUNCH OF Q-SYMPHONY BIOINFORMATICS COMPUTING SYSTEM: A HIGH-PERFORMANCE CLUSTER FOR ANALYSIS OF LARGE-SCALE GENOMIC DATASETS

A. Molkenov¹, A. Daniyarov¹, A. Sharip¹, A. Seisenova¹, D. Karabayev¹, U. Kairov¹

¹Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan

Keywords: bioinformatics, next-generation sequencing, whole genome analysis

Introduction: One whole human genome, provided by next generation sequencing platforms, in raw format takes 20 to 50 GB. In the course of bioinformatics analysis and data analysis, the data volume increases to 300-500 GB per genome. with an increase in the number of samples, the occupied volume increases. Such a large amount of data required for the analysis of whole genomes demands powerful computing power in the form of servers and data warehouses combined into clusters. We at Laboratory of Bioinformatics and Systems Biology have developed and launched Q-Symphony bioinformatics computing system called ("Qazaq Symphony of Bioinformatics") for bioinformatics analyses of solving large scale genomic datasets.

Materials and methods: The Q-Symphony bioinformatics computing system consists 12high-performance HPE servers: 1control node, 8 compute nodes, 1fat-memory compute node, and 2storage nodes. The system runs on Red Hat Enterprise Linux. The management node controls access to user profiles, data warehouse and Moab Workload Manager. The total number of processing cores is 172, the total amount of RAM is 3072GB, and the total storage capacity is 198 TB, a peak performance of the system of 7.3 TFlops. All nodes use high-speed Infiniband network connections, which allow the data exchange between nodes at 100 Gbps speed. The computational capabilities of the Q-symphony system allow us to evenly distribute resources for each task performed, monitor the load on processor and memory resources in real time, and queue and execute sequentially large lists of tasks.

Results: Benchmark measurements performed on Q-symphony system showed an increase of sub-tasks execution from 15 to 54 times compared to standard solutions built on similar computational processors.

Conclusion: The presence of Q-Symphony, well-established and proven bioinformatics methods will make it possible to successfully analyze large-scale human genomic data and determine structural genomic variants and carry out complex comparative and population analysis.