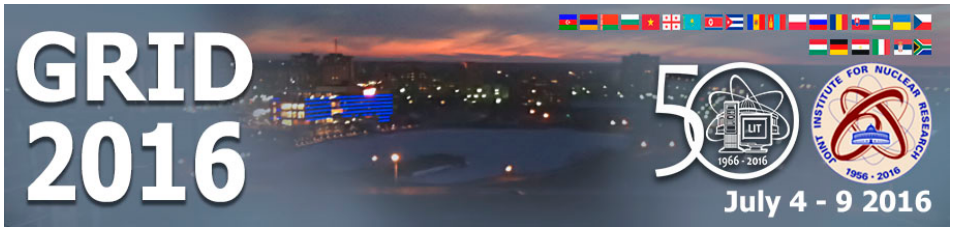

Contribution ID: **41** Type: **Sectional reports**

Optimization for Bioinformatics genome sequencing pipelines by means of HEP computing tools for Grid and Supercomputers

Tuesday, 5 July 2016 16:30 (15 minutes)

Modern biology uses complex algorithms and sophisticated software toolkits for genome sequencing studies, computations for which are impossible without access to powerful or significant computing resources. Recent advances of Next Generation Genome Sequencing (NGS) technology led to increasing volumes of sequencing data that need to be processed, analyzed and made available for bioinformaticians worldwide. Analysis of ancient genomes sequencing data using popular software pipeline PALEOMIX can require resource allocation of powerful standalone computer for a few weeks. PALEOMIX include typical set of software used to process NGS data including adapter trimming, read filtering, sequence alignment, genotyping and phylogenetic or metagenomic analysis. Organization the computation by sophisticated WMS and efficient usage of the supercomputers can greatly enhance this pipeline. Using related storage systems facilitate subsequent analysis.

Bioinformatics and other compute intensive sciences draw attention to the success of the projects which use PanDA beyond HEP and Grid. PanDA - Production and Distributed Analysis Workload Management System has been developed to address data processing and analysis challenges of ATLAS experiment at LHC. Recently PanDA has been extended to run HEP and beyond HEP scientific applications on Leadership Class Facilities and supercomputers.

In this paper we will describe the adaptation of the PALEOMIX pipeline to a distributed computing environment powered by PanDA for Ancient Mammoths DNA samples. We used PanDA to manage computational tasks on a multi-node parallel supercomputer. That was possible as we split input files into chunks which could be computed in parallel on different nodes as separate inputs for PALEOMIX and finally merge output result. We dramatically decreased the total computation time because of jobs brokering, submission and auto resubmission of failed ones by means of PanDA, which also demonstrated it earlier for the HEP applications in the Grid.

Thus using software tools developed initially for HEP and Grid can reduce computation time for bioinformatics tasks such as PALEOMIX pipeline for Ancient Mammoths DNA samples from weeks to days.

Primary author: Mr NOVIKOV, Alexander (National Research Centre "Kurchatov Institute")

Co-authors: Dr KLIMENTOV, Alexei (Brookhaven National Lab); Mr POYDA, Alexey (NRC KURCHATOV INSTITUTE); Mr TESLYUK, Anthony (NRC Kurchatov Institute); Mr NEDOLUZHKO, Artem (NRC Kurchatov Institute); Mr DRIZHUK, Daniel (NRC Kurchatov Institute); Mr SHARKO, Fedor (NRC Kurchatov Institute); Mr TERTYCHNYI, Ivan (NRC Kurchatov Institute); Mr MASHINISTOV, Ruslan (NRC Kurchatov Institute); Mr AULOV, Vasiliy (NRC Kurchatov Institute)

Presenter: Mr NOVIKOV, Alexander (National Research Centre "Kurchatov Institute")

Session Classification: 1. Technologies, architectures, models of distributed computing systems

Track Classification: 2. Operation, monitoring, optimization in distributed computing systems