Contribution ID: **112**   Type: **sectional reports**

High-throughput parallel pipelined data processing system for remote Earth sensing big data in the clouds

Wednesday, 2 July 2014 15:30 (30 minutes)

Report presents the approach and its software implementation of high-throughput parallel pipelined data processing system on the example of remote sensing of Earth from satellites (up to 1 Tb of data daily). The system is developed for processing potentially infinite data flow as it emerging in realtime. Data flow can be processing by set of subtasks steps(pipeline) with one or many required input files interconnected in a highly customizable way according to set of their parameters. That parameters could be outer or inner(metadata headers) file properties, for example sensor name, type, date, time, geolocation, besides file may have newer versions and be optional. New pipelines could be added, configured and tuned on the fly, sharing with priorities available cluster resources and new or the same data. The system operates at local common data storage for runtime computation and remote data sources and backup storage for results. Supported access and transport protocols for data at different locations programmingly extendable. Special description language flexsible adopts different computational tasks and applied solutions decreasing time for solution deployment and result gaining. Dynamically scalable by the system pool of cloud resources increases usage efficiency. The system controls pipelines, instances of subtasks and virtual machines lifetime, pipelines synchronization, statistics collection, errors detection and auto correction. The system is intended for sustainability with failover in any part of operation.

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

Co-authors: Mr POYDA, Alexey (NRC KURCHATOV INSTITUTE);   AULOV, Vasilij (National Research Centre ˝Kurchatov Institute˝)

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

Session Classification: Technology for storaging, searching and processing of Big Data

Track Classification: Section 3 - Technology for storaging, searching and processing of Big Data