# Machine learning technique for complex systems behavior prediction and anomaly detection for distributed data processing and management.

WLCG Demonstrator R&D Project Proposal
(draft v0.21, Mar 16, 2016)

Participants:

[1]Brookhaven National Laboratory - BNL (Upton NY, USA)
[2]European Particle Physics Laboratory – CERN (Geneva, Switzerland)
[3]Iowa University (Iowa, USA)
[4]National Research Center "Kurchatov Institute" - NRC-KI (Moscow, Russia)
[5]Tomsk Polytechnic University – TPU (Tomsk, Russia)
[6]University Texas at Arlington - UTA (Arlington TX, USA)
[7]Yandex School of Data Analysis (Moscow, Russia)
[8] Tomsk State University – TSU (Tomsk, Russia)

Leading PI : D.Duellmann[2] (CERN IT), A.Klimentov[1] (ATLAS), A.Ustyuzhanin[7](LHCb)
co-PI : M.Lassnig[2](ATLAS), S.Roiser[2] (LHCb)
Participants: J.Andreeva[2], F.Barreiro[6], M.Borodin[3,4], K.De[6],M.Grigorieva[4], M.Gubin[5], M.Hushchyn[7], S.Padolski[1], V.Parubets[5], I.Tertychnyj[4], A.Vaniachine[5,8]

We propose a one and half year (2016/06 – 2017/12) R&D project to evaluate how machine-learning techniques could be used for complex systems behavior prediction and anomaly detection. We will investigate the behavior of two vital systems for (at least) two LHC experiments: Data Management and Workload Management in Distributed Computing infrastructure. ~3 FTE will work on it from ATLAS and ~2 FTE from LHCb, the effort will be funded from grants received for R&D projects in Russia, USA and Europe.

The ATLAS and LHCb experiments at the LHC use the Worldwide LHC Computing Grid (WLCG) infrastructure for multi-petabyte data processing and analysis, marshaling a distributed, heterogeneous fabric of computing resources at collaborating institutes across the globe. The daily workload in the ATLAS experiment is about one million jobs, and up to 6PB of data are transferred weekly between WLCG sites.

WLCG infrastructure presents computing conditions in which contention for resources among high-priority data analyses happens routinely. Inevitably, over-utilized computing resources cause degradation of services or significant workload and data handling interruptions. For these and other reasons, LHC data management and processing must inevitably tolerate a continuous stream of failures, errors, and faults.

Since processing and transferring jobs generally complete or fail without producing side effects, the typical recovery mechanism simply involves automatic retry of failed jobs. Unfortunately, the simple approach results in an unpredictable delay in completion time dominated by repeated retries.

Our hypothesis is that application of statistical and machine-learning methodologies to the modeling of data processing and data management can guide the application of novel fault tolerance strategies with the potential to significantly reduce turnaround times for data intensive science at scale. There is a clear opportunity for a technological breakthrough, requiring innovative steps to enable significant improvements to turnaround times for exascale data processing, handling and analysis. Using a treasure trove of real-world data handling by PanDA/DIRAC/Rucio, we will investigate the nature of the failures and anomalies in the complex distributed computing infrastructure of WLCG, including networks. We will develop a data-driven model of short-lived correlated failures producing observed heavy-tail, non-Gaussian distributions in the evolving stream of random failures. We will apply time series modeling and online learning techniques to the continuous stream of information about processing failures and retries from the millions of jobs and data transfer requests in the ATLAS and LHCb distributed computing systems, in order to detect anomalies in failure rates, to identify clustered failures, and to determine failure characteristics that could support automated decision making regarding retry strategies. An example of such a failure characteristic is whether certain kinds of clustered failures are likely to be site- or node-specific, or rather indicative of issues that might occur more widely. Such statistical inferences can guide automated decision making, helping to determine whether it is worthwhile to rerun the scientific application on the same site or somewhere else, whether to choose a more sophisticated strategy, and whether strategies such as task molding and/or progressive job replication can be employed. Statistical inferences can inform development of online predictive models, and can further provide an early warning regarding whether human intervention is likely to be required.

During the first stage of the project we will identify and collect metadata that reflect functionality of various distributed computing entities including job schedulers, data transferring agents and analysis jobs. A common dictionary (ontology) has to be

defined to provide mapping of experiment-specific entities to a common space of notions and processes that would allow unification of collected data into a single scheme.

Storage with that scheme will be established with well-defined API to allow for storing metadata and querying it in an experiment-agnostic manner. Stored historical data should be anonymized in order to avoid risks of exposing experiment and/or user specific information. Metadata would serve as the basis for further research on a) various metrics that would reflect performance and anomalies of the distributed computing agents and infrastructure, and b) selection of different algorithms that would be able to predict anomalies or deviation of normal state of functioning. Initial research should be performed to provide guidelines and baselines for further investigations. The collected data and conducted research results will serve as the basis for further exploration of possible algorithms, policies that would minimize risk exposure by the network dysfunction and hardware failures. These datasets and metrics could serve as a wider-scale competition framework that could attract attention from a wider audience of Machine Learning experts for testing & improving predictive models. It will also be possible to add rules and meta-information into the content of a Data Knowledge Base or/and Catalog and Information system(s) and use them routinely in LHC data handling and processing.

During 2016-2017:

- Creation of experiment-agnostic ontology of distributed computing related entities
- Definition of mapping (dictionaries) between experiment-specific terms to that ontology (LHCb, ATLAS)
- Collection of anonymized historical data for experiments involved
- Identification of storage scheme for metadata (topology, state and log activity) and implementation of the API for storing / accessing the data
- A wide range of quality metric for anomaly detection / prediction will be compared and set of the most relevant will be selected
- A set of predictive models will be constructed and evaluated
- Several prospective models will be selected (prototype) and evaluated against chosen metric over collected data
- Specification of guidelines for creation and evaluation of new predictive model