Prominent ProdSys features

- Very high load;
- Changing structure;
- Size too large to feasibly simulate;

Some simulation frameworks (SimGrid) may be of use to simulate (small) parts of ProdSys, but simulating it entirely is currently not feasible.

ProdSys logs sources

- Oracle:
  - Production requests;
  - Dataset access logs;
  - Tasks;
  - Jobs.
- Hadoop:
  - Tasks;
  - Jobs;
- ElasticSearch:

Other

- Current ProdSys state according to monitoring;
- Date trends;
- Physic conference schedules;
- User input.

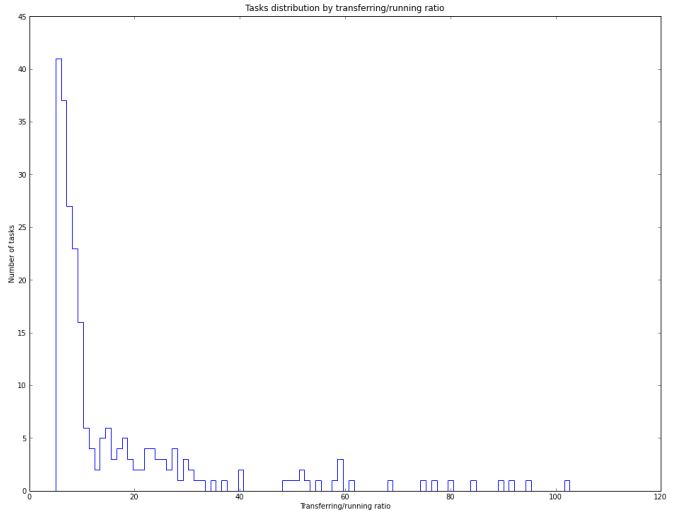
Anomaly detection proposal

ProdSys is too big to simulate directly;

ProdSys behavior is extensively logged;

ProdSys often exhibits anomalies:

- Unusual task completion times;
- Balancing issues;
- Connectivity issues;



It should be possible to model ProdSys behavior using Machine Learning-based models in conjunction with ProdSys logs and external data sources.

First step: “cold” prediction

- Based purely on ProdSys task execution logs;
- Does not make use of external data sources or ProdSys load;
- Based on historical averages.
- Considers all tasks of the same category (Project+Type+step) to have same duration.

Implemented.

Using ML for “hot” prediction

- “Hot” prediction will use current ProdSys state and external data sources as well as historical data;
- Data for “hot” predictions can be viewed as input vectors for ML-based models;
- There is enough historical data to teach ML models for prediction of ProdSys;
- The data is very diverse; preliminary data analysis (strong predictor detection) is underway.

Predicting ProdSys using ML-based models taught using historical data seems feasible, if data is prepared and pre-selected beforehand.