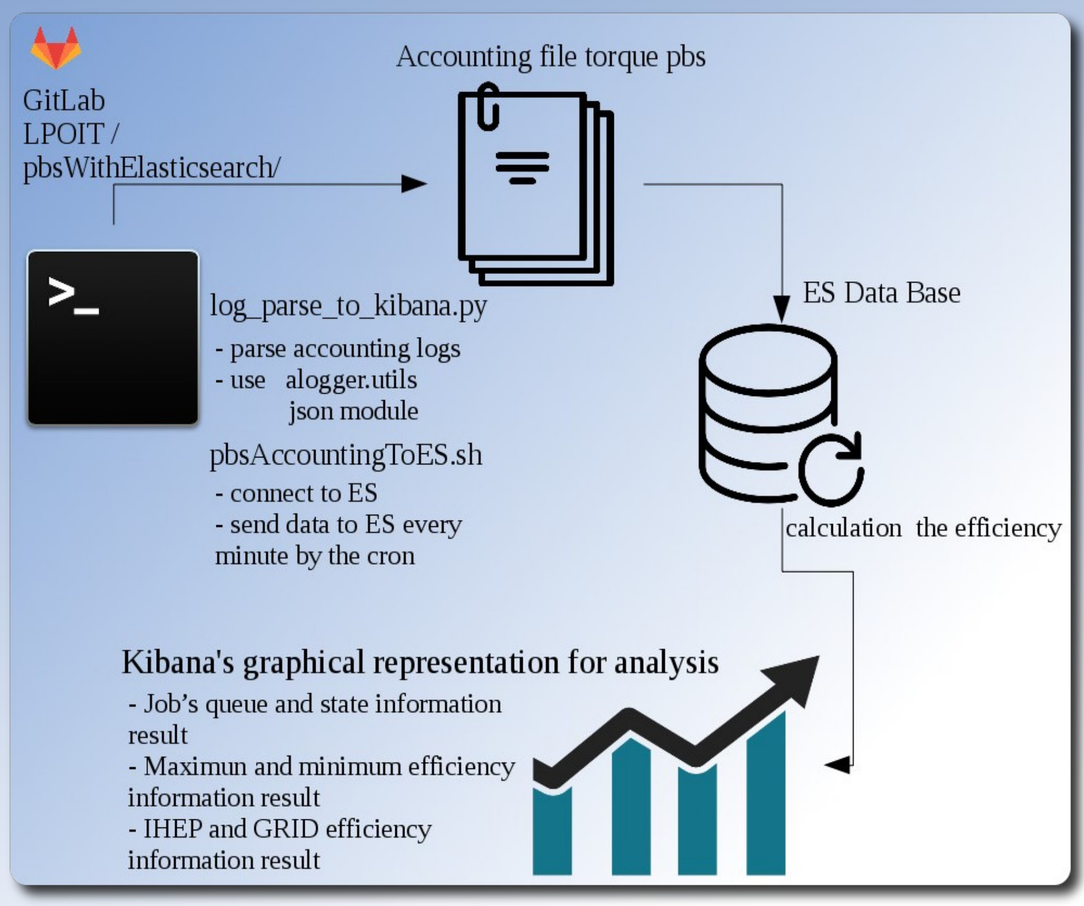
Efficiency measurement system for the computing cluster at IHEP

Ezhova Victoria, Viktor Kotliar

Institute for High Energy Physics named by A.A. Logunov of National Research Centre “Kurchatov Institute”,
Protvino, Russia

A lot of machine resources are expended on calculations related to research activities. We can estimate the size of the spent resources used for all types of tasks, make decisions for changing cluster configuration and to do the forecast for the work of the computer center in general. In this work you can see the calculations of the efficiency index and the graphical representation of work of a cluster on the basis of accounting information. It is one of the main tasks within work on creation of system of uniform monitoring of computer center of IHEP.

Review of the work performed and calculation of efficiency



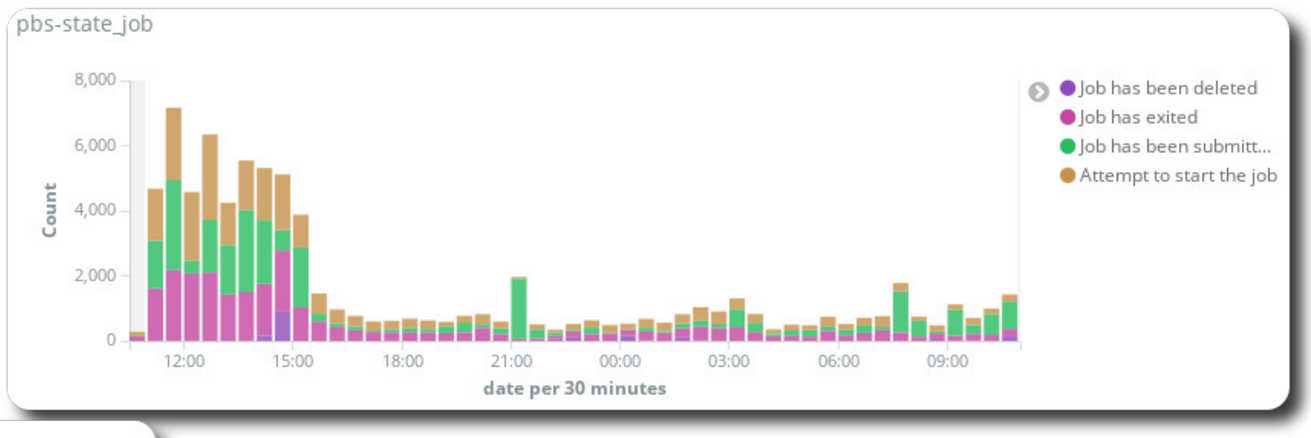
The python script sorts the accounting file of torque pbs into a python dict and throws off this information in JSON format. It use the alogger.utils - small python library to parse resource manager logs and json module which can generate JSON from python objects and lists. And then the bash script connects to Elasticsearch (ES) ant sents JSON data to ES every minute to display it in the form of the schedule from Kibana.

As next step Kibana calculates the efficiency indicator. It is the ratio of CPU time to walltime of a task which are given in the number format. The calculation also takes into account the presence or absence of ppn and ncpu indicators.

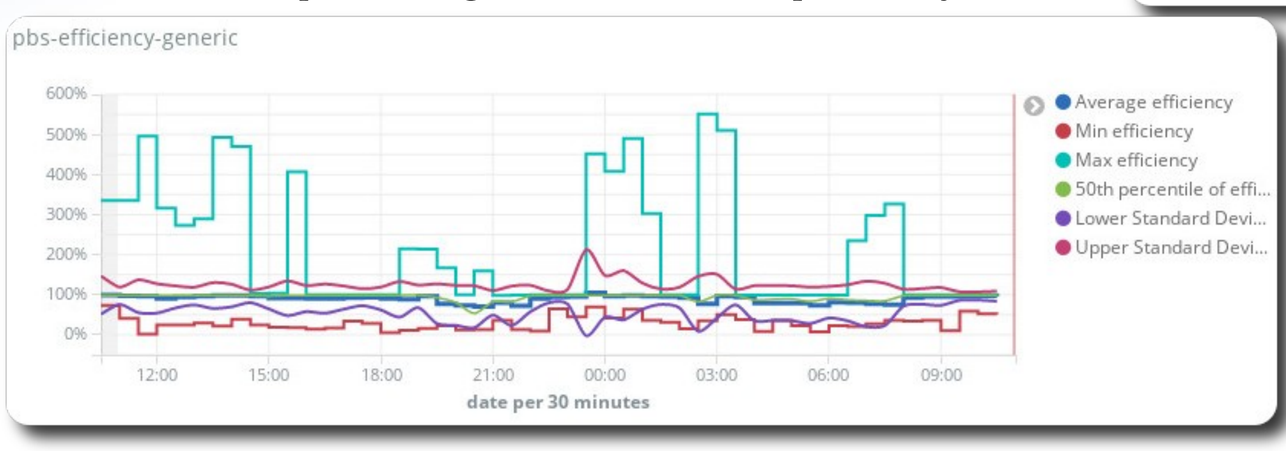
Script	Format
doc['cput'].empty ? 0 : doc['ppn'].value ? ((doc['cput'].value/doc['walltime'].value)/doc['ppn'].value) : doc['ncpus'].value ? ((doc['cput'].value/doc['walltime'].value)/doc['ncpus'].value) : (doc['cput'].value/doc['walltime'].value)	Percentage

Kibana’s graphical representation

Several graphs are constructed reflecting the growth or decrease in the efficiency of resource use by a group of tasks based on efficiency indicator.

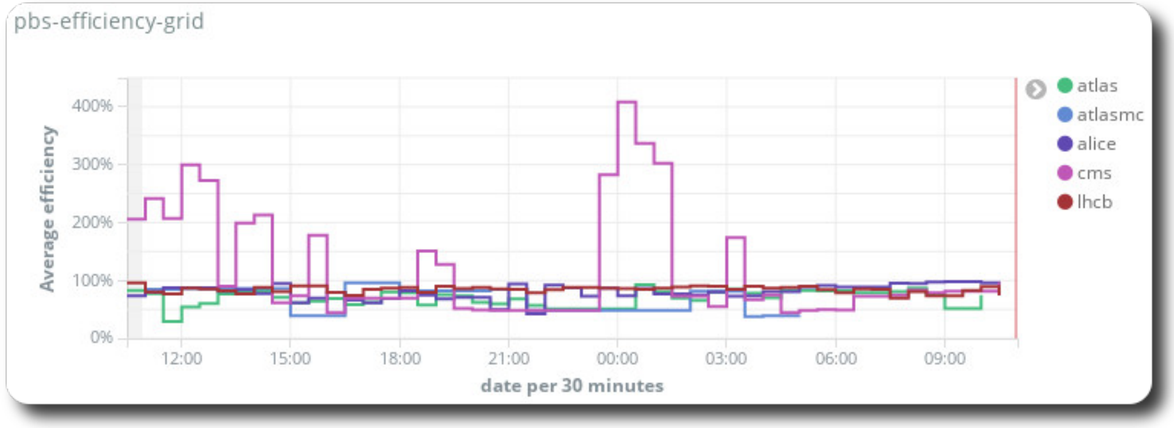


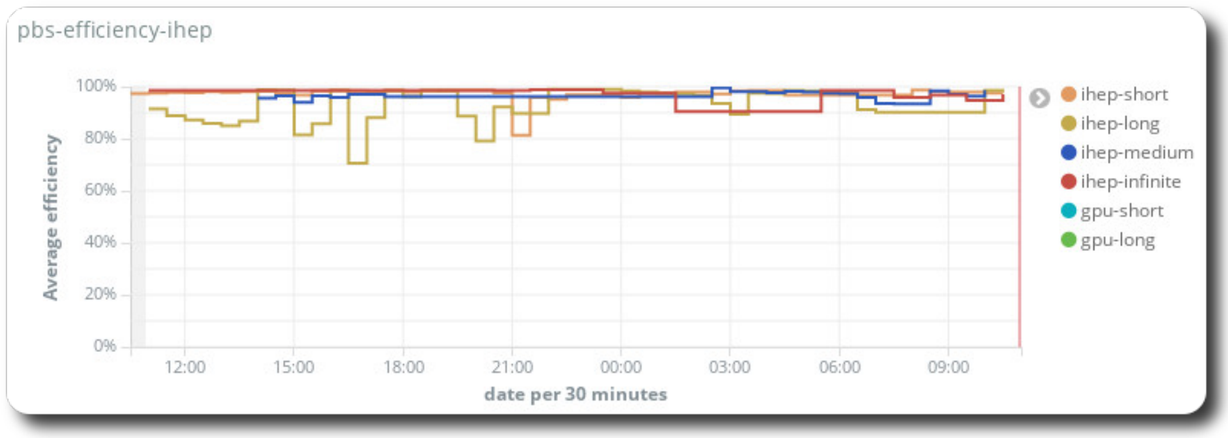
This is the histogram of the status of tasks whitch include the number of canceled, interrupted, completed tasks. You can see that the overwhelming number of tasks are completed tasks, those that are currently executing and those that are in the execution queue. Tasks were interrupted during the calculation are practically absent.

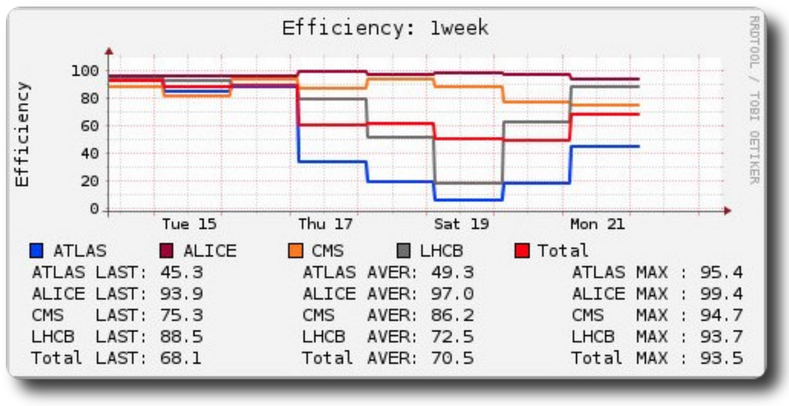


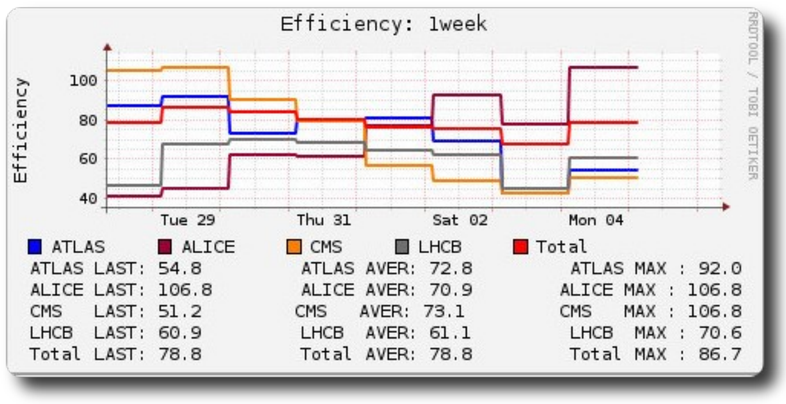
The linear graph reflects several critical indicators at once. With the example of maximum efficiency, you can observe abnormal deviations from the acceptable level. In fact, such tasks are bad, because the cputime of the task significantly exceeds the walltime to complete the task. In the future we will intend to implement a cluster management system that will evaluate the effectiveness to destroy bad tasks or decrease their number.

The last 2 graphs show the effectiveness of accomplishing the tasks of the IHEP users and grid tasks. The efficiency of tasks like Ihep-medium fell more often than others in the specified period period (one week). At the same time, there are not many such deviations, and on the whole the schedule looks more stable than the schedule for Grid tasks.









Memory allocation systems

By using developed efficiency mesurement system it was supposed to try different memory allocation systems on the computing nodes of the cluster and check their effect on the efficiency and memory usage. For these purposes tcmalloc and jemalloc packages were tested.

For an experiment the tcmalloc package was configured on the cluster working node for one week. During tcmalloc usage it is seen on the graph that a sharp decline of cpu usage occurred in compare with previouse gradual usage.

Jemalloc was tested for the same period. After inclusion of libjemalloc.so.1 library there was a growth of overall performance of a cluster (in particular CPU utilization). But processing jobs Failed - more than a half of jobs and about 27% of multy core jobs were interrupted.

That means tcmalloc and jemalloc are not usefull for the IHEP cluster.

Cluster Management System

In the future there is the plan to create additional control component for the Cluster Management System (CMS) with objectives to analyse efficiency indicators of overall cluster performance and to manage the cluster in a way of improve resorce usage efficiency.

At this stage CMS consists of event-driven management system, configuration management system, monitoring and accounting system and a chat-ops technology which is used for the administration tasks.