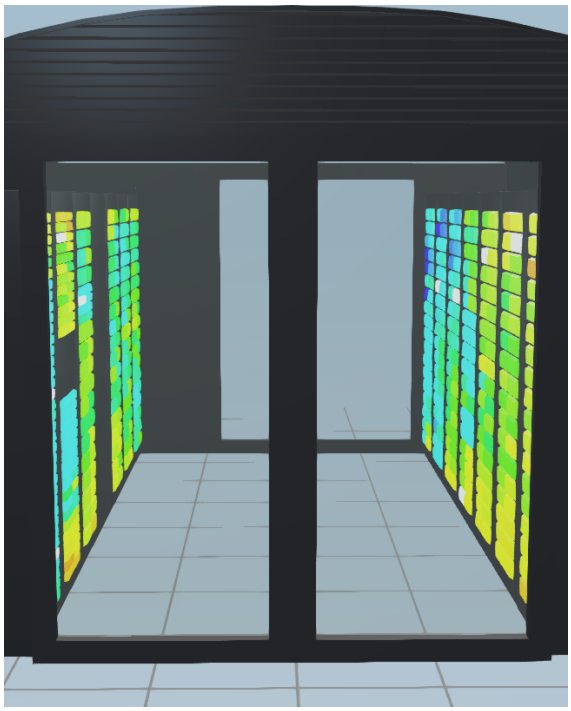
Visualising HPC System's Load



Petr Stehlík

Energy efficiency is one of the most timely problems in managing HPC facilities which can be addressed at different scale and perspective. Using Internet of Things technologies this project focuses on visualising data collected from the Galileo supercomputer in a web application.

The current monitoring system[1] consists of several layers which allow to aggregate in a single point heterogeneous data sources which consist of computing elements, node, job scheduler and facility telemetry of the Galileo supercomputer located at CINECA, Bologna, Italy.

The system was named *ExaMon*[1] and is built on top of MQTT protocol[2] which allows measured metrics to be send to a central broker where received data are processed and stored in KairosDB[3] database utilizing Cassandra cluster.

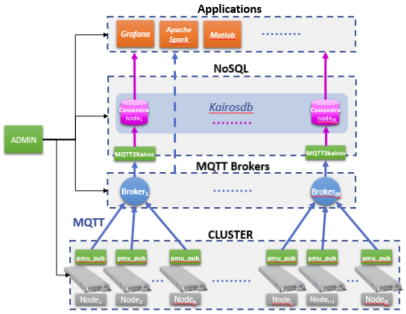


Figure 1: Examon Architecture

[1] Shorthand for Exascale Monitoring

This enables us to post-process data in time-oriented fashion in order to visualise them on a time-line and as a single number as well.

Current implementation uses Grafana framework to visualise data stored in KairosDB. Grafana will be replaced by the project's result of creating a dedicated web application for defined use-cases with 3D model of a cluster room showing various metrics of the whole HPC system with focus on energy consumption and efficiency.

Methods

The whole project can be separated into three phases. First phase is data anylysis where the whole dataset of available metrics was presented, how they are distributed and eventually processed on the back-end.

Data Analysis

Datasets can be divided into multiple levels of aggregation:

- Per-core level – the most low-level data can be found in core's registers such as IPS, Lx-cache misses and more. It also provides info about its load and temperature.
- Per-CPU level – each node consists of two CPUs and each of them can provide data about its C-states, energy counters and its frequency.
- Per-node level – most of the information available on node-level basis are coming from IPMI.[4] Via this interface we can access info about node's utilization, multiple temperature sensors and average power consumption.
- Per-cluster level – the Galileo's cluster room was equipped with several environmental sensors. This dataset is not currently available due to technical problems.
- Per-job level – data gathered using the PBS scheduler's hooks. This dataset is aside from previous ones since it points to allocated and used resources of the job submitted to the queue. This

data are stored directly to Cassandra cluster ommitting KairosDB.

With each level we can aggregate the lower levels (except job-level data). This is especially useful for core-level data which are mostly too dense for any comprehensible visualisation.

Visualisation

Second phase was to visualize data stored in KairosDB in simple yet insightful way in a lightweight web application. The application, called ExaMon Web, uses Angular framework as its base on top of which several other libraries were used. Worth mentioning are Dygraphs and Bootstrap. The former library produces powerful time-oriented charts utilizing the canvas element in web browser. The latter is a CSS framework to produce uniform user interface across the whole application.

Compared to Grafana, the created web application feels more lightweight, fast and easier to use because of the prepared datasets which are being used. The balance between configurability and the ease of use must have been found. We concluded the best way to achieve this was to enable time selection on given datasets but restrict configurability of the charts themselves. This way user is not bogged down with configuration and only focuses on prepared data.

If there is such desire to see other metrics the Grafana framework is still available right next to the ExaMon Web. As an additional feature, compared to Grafana, we can perform more advanced queries using the KairosDB REST API.

Live Data

The last phase was to utilize the live stream of MQTT messages right in the ExaMon Web. Two use-cases were defined for the MQTT messages depending on their origin.

Several job info MQTT messages are send during the job's lifecycle inside the PBS and every job is assigned a unique job ID. Using the ID user can subscribe to given messages and view various information on the ExaMon Web job dashboard. The dashboard also uses Cassandra cluster in case the job is already finished and stored in the cluster. This way user can see additional data about their job.

Using the job data a user can view detailed info about allocated resources of the given job such as CPU load, system utilization and more as seen in 2. With this information the user can asses some conclusions about their program. How effective it is, where are the slow parts and even perform a top-down analysis for performance issues. Also they can view how the program performed in terms of energy efficiency.

Second use-case is designed for general public and partly for system administrator. The use case is separated into two different parts. First one is very similar to job dashboards where data are displayed as time-serie charts with the difference in aggregation level which is at the cluster level. This means we can easily display, for example, the cluster's CPU load.

The other part is the most crucial in terms of interactive data visualisation. An accurate 3D model of the Galileo cluster was created using Blender and with the help of Blend4Web incorporated into the ExaMon Web. Even further, deeper integration was realized utilizing WebSockets (using Socket.io library) that enables us to create a reactive paradigm model instead of polling-based one. The model inside the page receives live data that has been published by the nodes and send to the broker. A subscription model was developed to accomodate large amount of visitors. The model then colours each node based on the minimum and maximum value of all received data. Weighted moving average was used in order to accomodate for sudden spikes in data using the given formula:

$$v_{new} = v_{current} + v_{previous} \times (1 - \alpha)$$

where $v_{previous}$ value is set to the first available value and $\alpha = 0.75$ as a default value was chosen based on short-term evaluation.

Results

The web application ExaMon Web is ready for production and is already running on one of CINECA's virtual machine in staging environment. ExaMon Web can be split to two major parts: 1. Tool for overseeing job submitted to PBS queue. 2. Cluster-level visualisation and analysis.

We will describe both parts of the ExaMon Web with a prepared use-case scenario which were created during development.

Job Visualiser

Cluster Visualiser

Discussion & Conclusion

Acknowledgements

(who helped me?) PRACE acknowledgement will be given together at the colophon. Site acknowledgement if required. Other acknowledgement if requested.

References

[1] Beneventi, Francesco, et al. "Continuous learning of HPC infrastructure models using big data analytics and in-memory processing tools." 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2017.

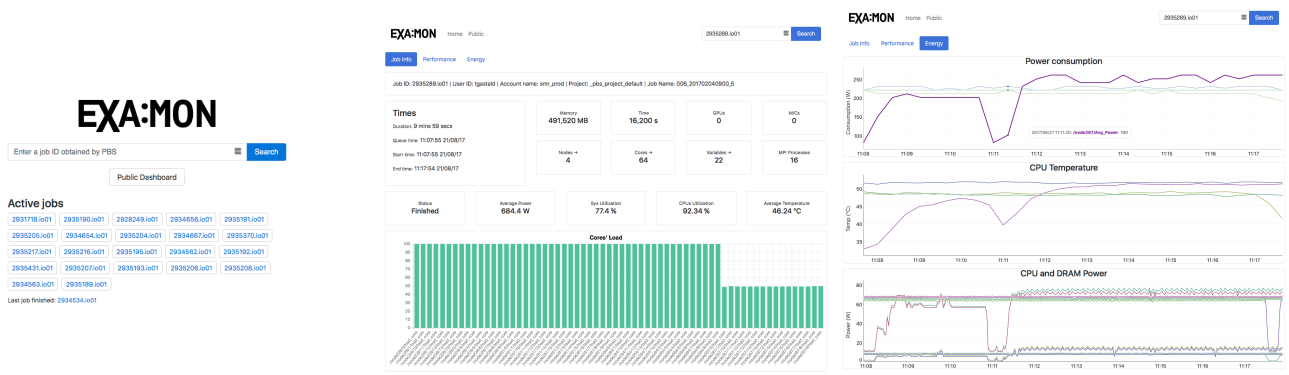
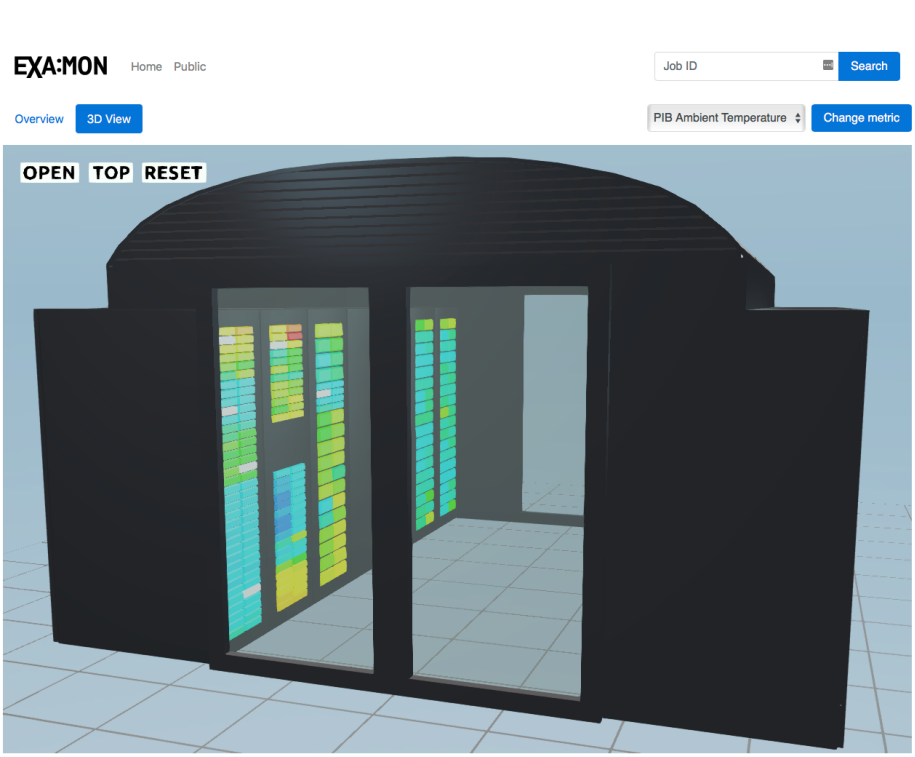
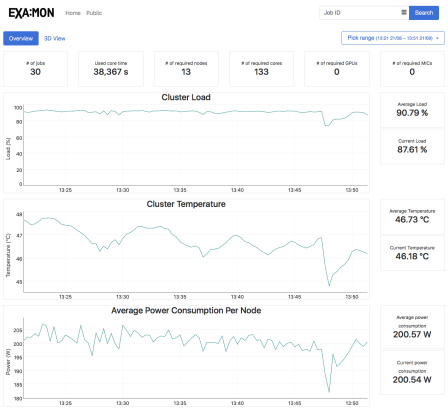
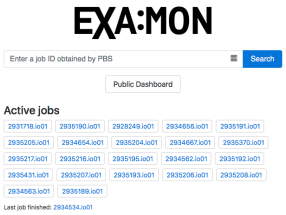


Figure 2: From left to right: intro page with jobs lookup, currently running jobs and last finished job; job's info dashboard with a finished job; job's energy dashboard.







3D model

[2] Locke, Dave. "Mq telemetry transport (mqtt) v3.1 protocol specification." IBM developerWorks Technical Library (2010).

[3] Goldschmidt, Thomas, et al. "Scalability and robustness of time-series databases for cloud-native monitoring of industrial processes." Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on. IEEE, 2014.

[4] Kaufman, Gerald J. "System and method for application programming interface for extended intelligent platform management." U.S. Patent No. 7,966,389. 21 Jun. 2011. APA

PRACE SoHPC Project Title

Web visualization of Energy load of an HPC system

PRACE SoHPC Site

CINECA, Italy

PRACE SoHPC Authors

Petr Stehlík, BUT, Czech Republic

PRACE SoHPC Mentor

Dr. Andrea Bartolini, UNIBO, Italy

Petr Stehlík

PRACE SoHPC More Information

Angular
Dygraphs
Bootstrap
Blender
Blend4Web

PRACE SoHPC Acknowledgement

Write any requested acknowledgements or thanks here. Mentors should be asked for them too.

PRACE SoHPC Project ID

1705