HAL
archives-ouvertes.fr

The Green Computing Observatory: a data curation approach for green IT

C. Germain-Renaud, F. Fürst, T. Jacob, M. Jouvin, G. Kassel, J. Nauroy, G. Philippon

▶ **To cite this version:**

C. Germain-Renaud, F. Fürst, T. Jacob, M. Jouvin, G. Kassel, et al.. The Green Computing Observatory: a data curation approach for green IT. First EGI Community Forum / EMI Second Technical Conference, Mar 2012, Garching, Germany. pp.059. in2p3-01015504

HAL Id: in2p3-01015504
http://hal.in2p3.fr/in2p3-01015504

Submitted on 26 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PROCEEDINGS OF SCIENCE

The Green Computing Observatory: a data curation approach for green IT

Cécile Germain-Renaud[a], **Frederic Fürst**[b], **Thibaut Jacob**[a], **Michel Jouvin**[c], **Gilles Kassel**[b], **Julien Nauroy***[d], **Guillaume Philippon**[c]

[a] *Université Paris Sud, Cecile.Germain@lri.fr, Thibaut.Jacob@u-psud.fr*
[b] *Université Picardie Jules Verne, {Frederic.Furst, Gilles.Kassel}@u-picardie.fr*
[c] *IN2P3-CNRS, {Jouvin, Philippo}@lal.in2p3.fr*
[d] *INRIA-Saclay, Julien.Nauroy@lri.fr*
Email: {Cecile.Germain, Julien.Nauroy}@lri.fr, {Frederic.Furst, Gilles.Kassel}@u-picardie.fr, {Jouvin, Philippo}@lal.in2p3.fr, Thibaut.Jacob@u-psud.fr

The Green Computing Observatory (GCO) is a collaborative effort to provide the scientific community with a comprehensive set of traces of energy consumption of a production cluster. These traces include the detailed monitoring of the hardware and software, as well as global site information such as the overall consumption and overall cooling. The acquired data is transformed into an XML format built from a specifically designed ontology and published through the Grid Observatory website.

EGI Community Forum 2012 / EMI Second Technical Conference,
26-30 March, 2012
Munich, Germany

*Speaker.

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike Licence.

PoS(EGICF12-EMITC2)059

1. Motivation and goals

The importance of energy saving in IT systems is so widely acknowledged that there is no need to detail it here, nor the explosion of related research; [1] presents a survey of the fundamental challenges and recent advances. [2] pointed some essential limitations on the path of energy-efficiency improvements as follows.

- Energy consumption is a complex system. Manufacturers have created sophisticated HW/SW adaptive control dedicated to energy saving in processors, motherboards, and operating systems, e.g. Advanced Configuration and Power Interface (ACPI), or the Intel technology for dynamically over-clocking single active cores. Administrators define management policies, such as scheduling computations and data localization with various optimization goals in mind. Finally, usage exhibits complex patterns too [3].

- The metrics remain to define. Energy efficiency should be the ratio of energy to service delivery, but for data centers and Clouds, service output is difficult to measure and varies among applications.

- Almost no public data are available, while benchmarking requires empirical data and ideally behavioral models.

The Green Computing Observatory (GCO) addresses the last issue within the framework of a production infrastructure dedicated to e-science, providing a unique facility for the Computer Science and Engineering community.

The first barrier to improved energy efficiency is the the lack of overall data collection on the energy consumption of individual components of data centers. The GCO collects monitoring data on energy consumption of a large computing center, and publishes them through the Grid Observatory portal www.grid-observatory.org. These data include the detailed monitoring of the processors and motherboards, as well as global site information, such as overall consumption and external temperature, as global optimization is a promising way of research [4]. A second barrier is making the collected data usable. The difficulty is to make the data readily consistent and complete, as well as understandable for further exploitation. For this purpose, the GCO opts for an ontological approach in order to rigorously define the semantics of the data (what is measured) and the context of their production (how are they acquired and/or calculated).

The GCO participates in the wider Grid Observatory initiative [5]. Its overall goal is to create a full-fledged *data curation* process, with its four components: establishing and developing a long-term repository of digital assets for current and future references, providing digital asset search and retrieval facilities to scientific communities through a gateway, tackling the good data creation and management issues, and prominently interoperability, through formal ontology building, and finally adding value to data by generating new sources of information and knowledge through both semantic and Machine Learning based inference. This paper reports on the first achievements, specifically the acquisition process, the ontology and its corresponding XML format.

PoS(EGICF12-EMITC2)059

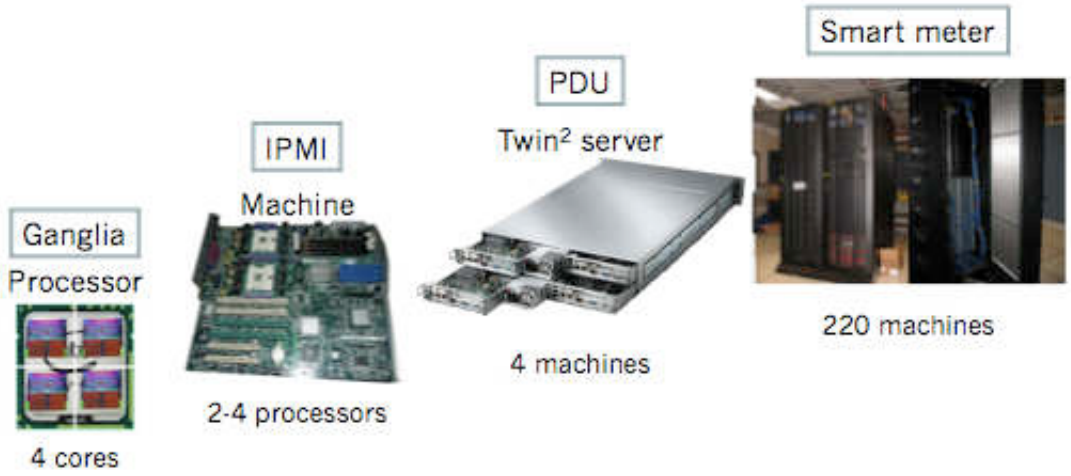

Figure 1: Monitoring instruments and scale

2. The acquisition apparatus

We define the data curation process by developing the complete hardware and software infrastructure for monitoring the computing center of Laboratoire de l'Accélérateur Linéaire (CC-LAL), and making the resulting data publicly available. The CC-LAL is mainly a Tier 2 in the EGI grid, but also includes local services and a Cloud infrastructure, with common characteristics of production quality and 24x7 availability. The computing site currently includes 13 racks hosting 1U systems, 4 lower-density racks (network, storage), resulting in ≈240 machines and 2200+ cores, and 500TB of storage. The following list shows why the CC-LAL is a good testbed.

- Heterogeneity, as the site features DELL and IBM systems, with classical and Twin^2 technologies. Thus, designing a general information model is mandatory.
- Classical, cold-water based central cooling (traditional racks plus cooling) as well as the more advanced water-cooled racks are present and monitored.
- The traffic is dominated by High Energy Physics experiments, which are high throughput, loosely coupled, data-intensive, as opposed to High Performance, compute intensive, strongly parallel workloads. We thus get some approximation of the behaviour of a data center.
- It hosts the experimental infrastructure of the StratusLab FP7 project, which builds and operates a Cloud on top of the EGI grid resources, thus creating an opportunity for Cloud-oriented monitoring.

Most work on energy consumption is based on the measurement of the inlets associated with the blades, through smart PDUs (Power Distribution Units). With the advent of Twin^2 servers, the granularity of energy consumption measurements becomes limited to the 8-16 processors of the server, which is clearly too coarse. We exploit the wealth of information provided by the IPMI (Intelligent Platform Management Interface) technology. IPMI is an industry agreement defining

PoS(EGICF12-EMITC2)059

a standardized, abstracted, message-based interface to intelligent platform management hardware, and standardized records for describing platform management devices and their characteristics. Besides power consumption, extremely detailed information about the motherboard, such as the operating voltage or the fan speed is available. Nonetheless, some servers are equipped with PDUs, in order to exploit the opportunity to calibrate one instrument with another. To be useful, the energy data have to be related to computational usage.

We use Ganglia to capture CPU, memory and network usage. A specific plugin is currently designed in order to report also on the ACPI states [6].

Finally, a smart meter reports on the overall energy consumption of the site, including central cooling, and its ambient temperature. Fig. 1 shows a typical configuration of the sources. With a 5 minutes sampling period, the volume is in the order of 1GByte per day.

3. The Information Model

To rigorously define the semantics of the data while addressing the problem of their heterogeneity, an ontological approach is implemented to define concerned entities and correlate them. This approach relies on an ontology of measurement providing a general framework [7] - *magnitudes* are measurements of *qualities* inhering in *objects* - which is refined into a model of energy consumption in computing centers. The general framework is itself an extension to the foundational ontology DOLCE [8]. *Objects*, considered in a broad sense, are physical ones (e.g. computers and associated components) or temporal ones such as *processes* (e.g. rotating movement of fan) or *events* bounded in time (e.g. motherboard failure). *Qualities* are dimensions of these objects which are observed and measured; they differ according to whether they are inherent to physical objects (e.g. temperature of a component), processes (e.g. speed of rotation of a fan) or events (e.g. duration of a power failure). Measurement instruments have their own qualities (e.g. resolution, calibration). *Magnitudes* can be boolean, numerical, scalars or vectors [9]. The magnitude (or measurement) assigned to a quality is obtained either by data acquisition from a sensor or by calculation from other qualities, for derived qualities (e.g. power is a function of voltage, intensity and power factor) or in the case of missing values (e.g. extrapolation).

4. Publication Format

This ontology translates into an XML schema. The schema involves metadata element, mainly reporting on the middleware, motherboards, sensors, and machines, and actual measurements in the form of time series. An overview of their relations is presented in Fig. 2 as an UML class diagram. The following elements are defined.

motherboard. This element describes the hardware elementary block of a computing system: the main board itself but also the CPU, memory and other hardware components. The motherboard is uniquely identifiable by its serial number, which is available through the IPMI protocol.

Each motherboard instance has a set of mutable properties, such as the firmware version or the chassis serial number - to help identify machines belonging to the same Twin^2 systems.

middleware. The middleware describes the software directly operating on top of a motherboard. Two types of middleware are present: regular Operating System (OS) when the node is

PoS(EGICF12-EMITC2)059

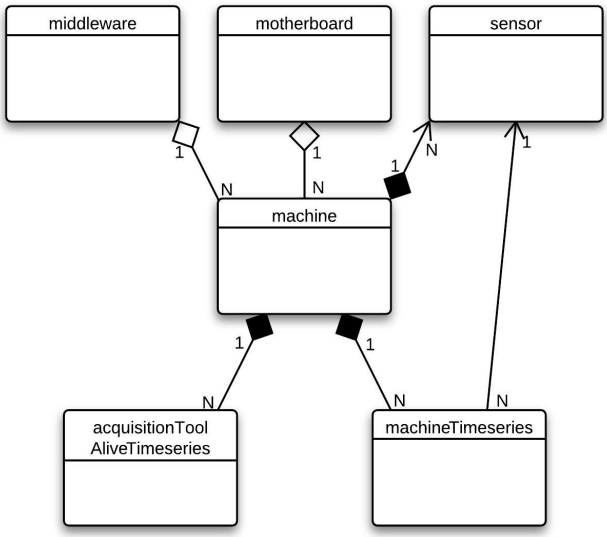

Figure 2: UML class diagram representing the main XML types

intended to support a non-virtualized workload, or a hypervisor in the virtual case. We chose this term in order to avoid the ambiguous term "OS", which can refer either to the hypervisor or to the software running inside a virtual machine. Each instance describes the general properties of the middleware, such as the kernel name and version. One instance of middleware can be associated with multiple machines at the same time. Installed or running software is not represented as middleware instances, as they are generic, but rather under the machine instances.

machine. We define this element as being the association between a motherboard and a middleware. Each machine is associated to a unique motherboard. On the contrary, a motherboard can be associated to multiple machines over time, but only to one at a given time.The association can change over time, when replacing the motherboard or upgrading the middleware; in both cases we consider a new machine has been created. This way we ensure that the interpretation of the acquired data is consistent over time: time series related to a given machine under the same firmware version can be considered as acquired under identical hardware and software conditions; thus the only variables are the external ones, such as computing workload and environment features. Conversely, different motherboards or OS may (and generally will) provide, not only different performance, but different monitoring data.

Each machine instance has a set of mutable properties, such as the IP address, network name and domain. These properties can be modified over time without modifying the semantics of the acquired values.

sensor. Each machine comprises a set of sensors, which can be either physical or software. Values are acquired from these sensors through an acquisition tool and protocol. In our case, these tools are Ganglia, IPMI and PDU over SNMP. Each sensor can hold a set of information regarding its acquisition capabilities, such as the accuracy, precision, sensivity or measurement range, as described in the SSN Ontology [10]. This information is sometimes directly provided by the sensors themselves.

machineTimeseries. This element represents a time series. Each instance refers to a machine

PoS(EGICF12-EMITC2)059

and a sensor. By convention, one day of acquisition is stored in each time series instance. In our model, time series do not directly refer to the hardware or software part to force the evaluation of the acquisition context: as an example, the power consumption, measured at the hardware level, cannot be analyzed without taking into account the CPU usage, measured at the software level. With our representation, both power consumption and CPU usage time series refer to the same machine and can easily be interpreted as two facets of the same phenomenon.

acquisitionToolAliveTimeseries. This element stores the availability over time of an acquisition tool for a given machine. This information is stored as a series of transitions (down→up or up→down) to save space. As single values or entire timeseries could be missing in a given data set, knowing if the acquisition tool was available will help understanding whether the machine was not monitored or the acquisition system was undergoing a transient failure.

Implementation further requires supporting scalable exploration (selection, projection, and more complex requests) of the datasets. Given the performance limitations of XML querying, we decided to offer maximal flexibility to the user. Firstly, a common schema is used for the three acquisition sources (IPMI, Ganglia and PDU), and files are structured at a fine grain (one per day, machine and source) ; flexible aggregation is made possible through the standard XInclude tool. Second, we publish both the native bulky data, but also selected ones, with a typical 70% volume reduction. One typical week of published data still represents 3GB in XML format.

The next step is to integrate this approach into a higher level view, including both acquisition from other sites, and the data dissemination process. It should be oriented towards users and usage, statistical analysis of time series. SDMX (Statistical Data and Metadata Exchange) [11] is a de-facto standard (and ISO norm for SDMX1.0) within the sphere of economic data and the extension of the ontology will be done in line with the SDMX information model.

Finally, the issue of Linked Data [12] should be considered: Examples such as [13] indicate that building on the SDMX experience makes the transition to Linked Data access manageable.

5. Data Visualization

A custom visualization tool has been created to unify the visualization of the different data sources. We use a modified version of the Ganglia web visualizer, which is one of the most popular monitoring tools for systems administrators. We modified it to replace its real-time monitoring functionality, which implies loosing precision for old data, with the ability to load weekly archives of past acquisitions at their original frequency.

Fig. 3 shows how these different sources can be visualized together and used as a first approach for understanding the relationship between CPU usage, power consumption and temperature. The graphs are the following:

- top left, the CPU activity, in percentage, as returned by Ganglia,
- bottom left, the average power, in Watts, as returned by IPMI,
- top right, the internal temperatures, in degrees Celsius, as returned by IPMI, with the Ambient temperature being correctly acquired, the MCH temperature being visibly faulty, and the CPU temperatures completely missing,

PoS(EGICF12-EMITC2)059

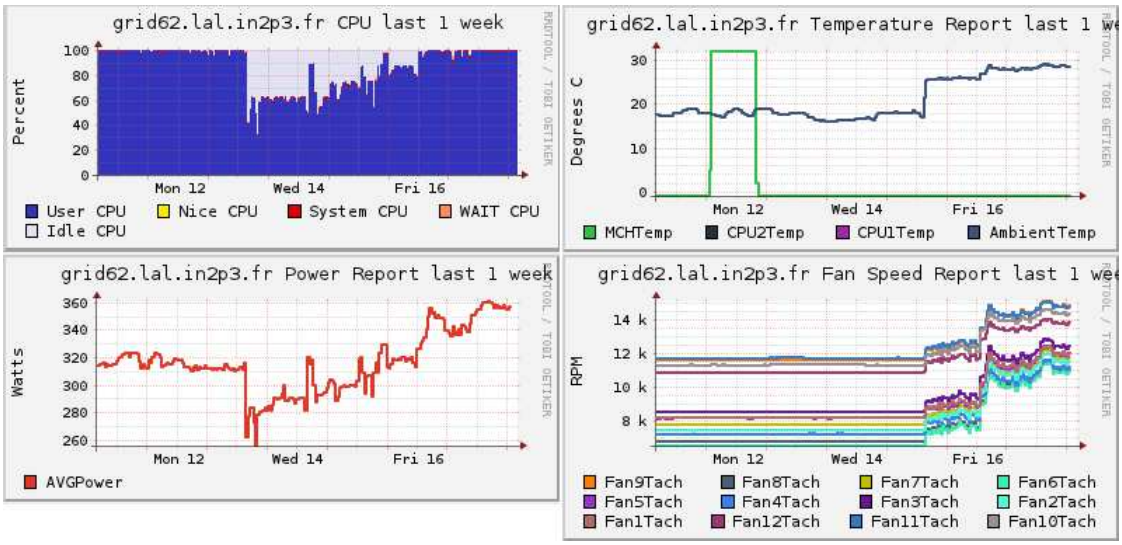

Figure 3: Visualizing together data from multiple sources

- bottom right, the fans speed, in Revolutions per Minute (RPM), as returned by IPMI.

One can see that the CPU is used at 100% at both the beginning and the end of the week. This translates into two power peaks at or above 320 Watts. However, the end of the week power peak is much higher that the one at the beginning (360W versus 320W). The CPU usage is identical, and the explanation is elsewhere, namely by the fans speed being much higher on friday that on monday; this, in turn, can be explained by the brutal step in temperature, from 18 degrees Celsius to nearly 30. The visualization makes evident this change of regime, due to external factors, and not the workload. It may considerably help eliminating outliers prior to analysis, or conversely, identifying exceptional conditions (such those prevailing at the end of this week) leading to extreme events.

This example shows the importance of integrating the data acquired at the hardware and software levels. It also shows that the models and simulations of energy usage must take into account external factors such as the ambient temperature, and a real-world monitoring repository such as the GCO as providing a highly valuable source of experimental data covering a wide range of behaviors.

6. Conclusion

The final motivation of monitoring and semantics is optimization. The Green 500 initiative has pioneered an energy benchmarking approach for supercomputers, based on the widely accepted Top 500 benchmark; benchmarks oriented towards Cloud computing workloads have started to emerge [14]; however, the applicability of the benchmarking approach might be quite limited for large scale distributed systems, as real world experimentation is hardly possible on production system, e.g. for scheduling. Experimenting with simulators, as a fallback require large data sets. The GCO provides such data sets.

PoS(EGICF12-EMITC2)059

Moreover, an alternative to experimenting on real, large, and complex data is to look for well-founded and parsimonious representations and generative models from the large dimension space available from the detailed monitoring. Autonomics calls for models too: for real use cases, the knowledge at the core of the MAPE-K loop (monitor-analyze-plan-execute) has to include an off-line component, which cannot be purely a-priori, but must be built from historical data. The generative model approach has started to gain acceptance in the distributed system research (for two recent examples with special interest in non-stationarity, see [15] and [16]); the GCO data will contribute to dimension and validate models of energy usage in the same way.

Acknowledgments

This work has been partially supported by the EU FP7 project EGI-InsPIRE INFSO-RI-261323, the France Grilles initiative, and the CNRS program PEPS 2011-2012.

References

[1] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," *Advances in Computers*, pp. 47–111, 2011, marvin V. Zelkowitz (editor).

[2] U.S. Environmental Protection Agency, "Report to Congress on Server and Data Center Energy Efficiency," Tech. Rep., 2007.

[3] I. Rodero and al., "Energy-efficient application-aware online provisioning for virtualized clouds and data centers," in *Int. Conf. on Green Computing*, 2010, pp. 31–45.

[4] R. Das, J. O. Kephart, J. Lenchner, and H. Hamann, "Utility-function-driven energy-efficient cooling in data centers," in *7th Int. Conf. on Autonomic computing*, 2010, pp. 61–70.

[5] C. Germain-Renaud and al., "The Grid Observatory," in *11th IEEE/ACM Int. Symp. on Cluster Computing and the Grid*, 2011, pp. 114–123.

[6] *The ACPI Specification - Revision 5.0*, Advanced Configuration & Power Interface Std., 2011, http://www.acpi.info/.

[7] W. Kuhn, "A functional ontology of observation and measurement," in *3rd Int. Conf. on GeoSpatial Semantics*, 2009, pp. 26–43.

[8] S. Borgo and C. Masolo, "Foundational choices in dolce," in *Handbook on Ontologies*, 2nd ed. Springer, 2009, pp. 361–362.

[9] T. Gruber and G. Olsen, "An ontology for engineering mathematics," in *4th Int. Conf. on Principles of Knowledge Representation and Reasoning*, 1994, pp. 166–181.

[10] *Semantic Sensor Network XG Final Report*, W3C Incubator Std., 2011.

[11] *SDMX2.1 Technical Specification - Section 2: The Information model*, The SDMX initiative Std., 2011.

[12] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *Int. Jal. on Semantic Web and Information Systems*, no. 3, pp. 1–22, 2009.

[13] R. Cyganiak and al., " Semantic Statistics: Bringing Together SDMX and SCOVO," in *Linked Data on the Web Workshop at the 19th Int. World Wide Web Conference*, 2010.

PoS(EGICF12-EMITC2)059

[14] M. e. a. Ferdman, "Clearing the clouds: a study of emerging scale-out workloads on modern hardware," in *Proc. 17th int. conf. on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '12. ACM, 2012, pp. 37–48.

[15] H. M. N. D. Bandara and A. P. Jayasumana, "On characteristics and modeling of p2p resources with correlated static and dynamic attributes," in *GLOBECOM*, 2011, pp. 1–6.

[16] T. Elteto, C. Germain-Renaud, P. Bondon, and M. Sebag, "Towards Non-Stationary Grid Models," *Journal of Grid Computing*, no. 4, Dec. 2011.

PoS(EGICF12-EMITC2)059