Monitoring Power Data: A first step towards a unified energy efficiency evaluation toolset for HPC data centers

Hayk Shoukourian, Torsten Wilde, Axel Auweter, Arndt Bode

Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities, Bolzmannstraße 1, 85748 Garching bei München, Germany

Technische Universität München (TUM), Fakultät für Informatik I10, Boltzmannstraße 3, 85748 Garching bei München, Germany

Article history:
Received 24 May 2013
Received in revised form 25 November 2013
Accepted 28 November 2013
Available online xxx

ABSTRACT

The energy consumption of High Performance Computing (HPC) systems, which are the key technology for many modern computation-intensive applications, is rapidly increasing in parallel with their performance improvements. This increase leads HPC data centers to focus on three major challenges: the reduction of overall environmental impacts, which is driven by policy makers; the reduction of operating costs, which are increasing due to rising system density and electrical energy costs; and the 20 MW power consumption boundary for Exascale computing systems, which represent the next thousandfold increase in computing capability beyond the currently existing petascale systems. Energy efficiency improvements will play a major part in addressing these challenges.

This paper presents a toolset, called Power Data Aggregation Monitor (PowerDAM), which collects and evaluates data from all aspects of the HPC data center (e.g. environmental information, site infrastructure, information technology systems, resource management systems, and applications). The aim of PowerDAM is not to improve the HPC data center’s energy efficiency, but is to collect energy relevant data for analysis without which energy efficiency improvements would be non-trivial and incomplete. Thus, PowerDAM represents a first step towards a truly unified energy efficiency evaluation toolset needed for improving the overall energy efficiency of HPC data centers.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Several studies show that over the past decade the worldwide energy consumption of data centers is increasing (Koomey, 2008, 2011). If this trend continues, the energy costs of future systems will start to be a dominant factor for the total cost of ownership (TCO) defined as the entire costs spent on using and acquiring assets (Wouters et al., 2005). This issue can already be seen at the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences where the predicted electrical energy costs over the lifetime of the newest HPC system approaches one half of the investment costs, whereas, the electrical energy costs of the earlier supercomputers at LRZ comprised only 1/10 of the investments costs in the timeframe of 2000–2005 and 1/5 in the timeframe of 2005–2010. Moreover, energy consumption is becoming a critical design constraint for future HPC systems. One of the major challenges in the race of building an Exascale supercomputer, capable of executing $10^{18}$ floating point operations per second (FLOPS), is energy consumption (Kogge et al., 2008).

In order to improve the energy efficiency in a data center the following items are necessary: a) collected data from all aspects of the data center (e.g. environmental information, site infrastructure, information technology systems, resource management systems, and applications); b) the ability to correlate data in order to better understand the interactions between different components of the data center, to assess the status of current key performance indicators (KPI), to identify improvement areas, and to verify the success of the optimization; and c) the provision of KPI information to data center operators and policy makers. While several techniques for improving the energy efficiency of data centers are known, there is currently a lack of a unified monitoring and evaluating toolset (Auweter et al., 2011) that addresses item a), and subsequently also items b) and c). A unified energy monitoring toolset is crucial for a holistic evaluation of the HPC data centers. Since data centers are unique in terms of (i) used monitoring and management tools, and (ii) residing HPC systems (system
hardware, system management software, etc.) the mentioned unified energy monitoring toolset should be independent from the existing data center specific monitoring and analysis tools. Without this wholistic view it would be non-trivial to achieve the optimal energy efficiency in the HPC data center. The importance of such a monitoring toolset can be summarized in the quote from management consultant Peter Drucker — “if you can’t measure it, you can’t manage it” (qtd. in Leyland (2009)).

Despite the presence of various KPIs that aim to assess the energy efficiency of HPC systems, most of them do not take into account the effects of the scientific applications running on those systems. For example, the FLOPS/watt KPI, used by Green500 (The Green500, 2013), does not completely reveal the overall energy efficiency of HPC systems, since it only focuses on compute bound scientific applications and does not consider the energy costs of data movement, either in memory or over the interconnect network (Bekas and Curioni, 2010; Henneke et al., 2012). Hence, there is a need for a new metric which will reflect the amount of energy required to solve a specific application with a certain input data set on a given hardware platform. Energy-to-Solution (EtS) (Minartz et al., 2011; Auweter et al., 2011) is one of such metrics that PowerDAM reports on. This metric indicates the aggregated energy consumption of a given application including compute nodes and parts-in-sub-system components (as system networking, system cooling and infrastructure) which were utilized during the application execution phase. EtS can be further used for:

- power capping

  The power behavior of applications will allow to evaluate the required cooling capacity of the system, to control specific power consumption constraints with utility provider, and to stay within the desired energy budget;

- understanding the power profile of executed applications

  Most of the HPC users (e.g. groups from industrial, research and educational institutions, etc.) can be charged according to consumed CPU hours. The knowledge about the energy consumption of applications will increase the users awareness of power consumption and will allow to have energy-driven charging policies.

This paper addresses the first step towards a unified energy measuring and evaluation toolset which is capable of monitoring and analyzing the energy consumption of a supercomputing site in a wholistic way, combining the HPC systems with data from the cooling, building infrastructure, and HPC applications. The developed toolset, called Power Data Aggregation Monitor (PowerDAM), was specifically designed to collect data from all aspects of a data center and to be independent of specific vendor hardware or existing data center tools. At LRZ, PowerDAM has been set up to monitor two HPC systems as well as their respective resource management systems (also called batch scheduling systems). PowerDAM is capable of monitoring not only HPC systems but any other system that can be represented in a hierarchical tree structure (e.g. the building infrastructure). It is able to monitor physical sensors as well as virtual sensors which can represent different functional compositions of several physical sensors.

The rest of this paper is organized as follows. Section 2 presents the background information and outlines the state of the art of the energy efficiency domain as important to LRZ. Section 3 introduces PowerDAM and describes its workflow and framework. This section also outlines the actual setup of the developed tool at LRZ. Section 4 discusses the aggregated energy consumption calculation of a given application (referred to as EtS) and presents the current functionality of PowerDAM via an example application on two different HPC systems. Section 5 delineates the visualization and reporting capabilities of PowerDAM. Section 6 describes the future development directions of PowerDAM and finally, Section 7 concludes the paper.

2. Background and problem statement

The increase of the compute density of modern HPC systems accompanied with rising electrical energy costs, and data center sustainability have become a major concern for policy makers, data center operators, and researchers. Several standardization bodies are addressing issues related to the efficient energy management of data centers and supercomputers. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) (American Society of Heating, 2013), is developing and promoting guidelines for efficient data center operations. Another institution, which is developing a common set of data center energy efficient key performance indicators (KPIs), is The Green Grid (The Green Grid, 2013). According to The Green Grid, the efficiency of the energy consumption spent on the HPC system site infrastructure can be quantified using the Power Usage Efficiency (PUE) metric defined as the ratio of the total energy used by the data center site divided by the amount of energy that is consumed by all of the IT equipment in the data center. Thus, for the detailed energy consumption analysis of the entire HPC system, it is important to not only monitor the compute nodes, interconnect network, and storage devices, but also the internal infrastructure such as cooling and power components (Auweter et al., 2011). This will further help to generate, verify, and validate model based descriptions of energy flows and data center thermal topology, whose necessity is also emphasized in the “Data Center Cooling Management And Analysis — A Model Based Approach” article (Zhou et al., 2012).

Current building infrastructure management systems, present in modern data centers, control and monitor infrastructure related components. This infrastructure control and management system is completely isolated from the behavior of the HPC systems residing in the data center and does not provide a control-side coupling between building infrastructure and HPC system management. Therefore, it is typically optimized for the operation of IT systems at maximum power draw.

Fig. 1 shows the complete facility power consumption profile of the LRZ data center (Complete Facility Power Profile), the latest HPC system (SuperMUC Power Profile), and the Linux cluster (Linux Cluster Power Profile) for a period of 8 days in December 2012. Contrary to the nearly constant power consumption of the Linux cluster, which is similar to the older generation of supercomputers at LRZ, SuperMUC shows a variation of up to 800 kW and dominates the overall power profile of LRZ.

A maximum power draw of 3.6 MW was observed during the Top500 (Top500, 2013) Linpack benchmark on SuperMUC. Therefore, it is possible to have a 2.7 MW variation in power consumption of SuperMUC during normal operation. One consequence of this highly dynamic power profile is a strong fluctuation in the building infrastructure efficiency. Only by correlating the infrastructure data with HPC system data is it possible to identify and address the inefficiencies present in the installation. This correlation will provide a wholistic view across multiple sources and structures and will allow a better understanding of the complete data center power profile. This understanding will lead to better optimization techniques which are necessary for reducing the overall energy consumption and the total cost of ownership (TCO) of the data center.

According to principles of energy efficiency in HPC (Auweter et al., 2011), the energy monitoring infrastructure should be able, in addition to collecting information about the system’s workload, to:...
and various utilization metrics (e.g. CPU load, temperature per core, per-node information on the interconnect network load, etc.), to collect information regarding all the applications that are running on the given system. This will allow for the measurement of the amount of consumed energy for a given application. It is also mentioned in Auweter et al. (2011) that there is a lack of such a unified measuring and monitoring infrastructure despite the presence of necessary individual components.

A set of related works on energy profiling infrastructure is found in Da Costa et al. (2009) and Ge et al. (2010). The first one presents a framework for energy efficient management of large scale distributed systems which does not consider the infrastructure relevant data. The latter one proposes a measurement evaluation tool as a combination of software components and additional external hardware components (e.g. digital meters), meaning that software can only be used with custom hardware. However, this approach cannot be applied to highly integrated large-scale systems since it would require a complete instrumentation of the entire supercomputer which are already starting to have more than 10,000 nodes. That is why more and more vendors aim to provide the measuring infrastructure inside the actual system. One of the institutions which considers this issue and defines measuring capabilities for new HPC systems is the Energy Efficient High Performance Computing Working Group (EE HPC Working Group) (Energy Efficient HPC Working Group, 2013). EE HPC Working Group is also concerned with defining metrics for the evaluation of HPC systems and data center efficiency. They develop guidelines for what types of measurements need to be taken, where in the data center they are to be taken, and the method in which the measurements are taken.

Another set of related tools in the area of operational management are Rackwise and openDCIM (openDCIM — Web Based Data Center Infrastructure Management Application, 2012; Rackwise Data Center Software, 2011), which are examples of data center infrastructure management tools (DCIM). These tools were developed to monitor and analyze the physical assets and resources within a data center. However, they do not consider the system software side and do not provide complete information on energy consumption of all components which are present in the compute system. Without the information on energy consumption of running applications the complete exploitation of the energy saving features provided by the target platform would be relatively complicated. The information on energy consumption of a given application will allow for the further understanding and tuning of the application internally (via change of algorithms, memory access patterns, etc.) as well as externally through hardware adaptation (e.g. static/dynamic voltage frequency scaling).

One could argue that PowerDAM could have been implemented as an extension to one of these DCIM tools in order to supplement the lacking functionalities. But in this case, the usage of PowerDAM would have been restricted to data centers that use a certain DCIM which was not the purpose of PowerDAM. The main aim of PowerDAM is to collect and correlate data from all aspects of a data center, where each of these aspects can have different hardware and/or management systems. This scenario is present at LRZ, where each of the data center aspects (i.e. building infrastructure, HPC systems, etc.) has different management tools.

In summary, within the problem domain of energy efficiency the aforementioned approaches are solution based (i.e. they pose a specific question and develop their system in order to answer that question). Whereas, PowerDAM looks generically at the problem domain. It gathers data in order to answer questions that may arise within the problem domain with little to any further development. PowerDAM was specifically developed to address the lack of:

- a unified power measuring and monitoring infrastructure;
- a tool that collects data required for the evaluation of energy efficiency data center KPIs;
- a tool that can correlate energy-data related to infrastructure and HPC systems;
- the ability to assess the energy consumption of specific applications; and
- a tool which gathers data that allows data center operators and policy makers to answer any question related to improving the energy efficiency of a data center.

This energy monitoring toolset is essential in gathering and analyzing power usage data for a data center and its HPC systems. The collected data is necessary for: investigating the power consumption of applications and new algorithms; assessing the energy saving features of hardware platforms; supporting the improvement of the data center infrastructure efficiency; helping certain
workload managers in energy efficient decisions; and assisting the fine-tuning of developed models, metrics, and intelligent process controls. All of these will further contribute to the overall sustainability of the data center.

3. Overview of PowerDAM

Power Data Aggregation Monitor (PowerDAM), which is under development at LRZ, is a unified energy measuring and evaluating toolset for HPC data centers. PowerDAM covers a wide range of power consumption analysis capabilities from the complete data center view (e.g. ability to report on PUE ([The Green Grid, 2013])) over the individual systems (e.g. ability to report on power consumption of compute nodes) to the user applications (e.g. ability to report on EtS). It takes as input the data fetched from sensors and resource management tools and outputs the information on energy efficiency KPIs (Fig. 2). PowerDAM uses an agent based data communication model for actual data retrieval, in other words, it maintains scripts/daemons which reside on the monitored entity/system side and have the permission to access and to push the requested data over a network to PowerDAM, where it is stored. This approach makes PowerDAM loosely coupled with the systems from which data is collected. This allows for relatively easy extension of the monitored system set.

The main design goals for PowerDAM were:

- **independence of monitored systems**

  Different HPC systems are manufactured by different hardware vendors (Intel, IBM, HP, etc.), and, therefore, the way hardware specific data (as power, load, temperature, etc.) is accessed can vary from system to system. The tool is not tied to a specific hardware, and, therefore vendor;

- **extensibility of monitored sensor/system set**

  Because of the variety of systems present in the data center an easy integration of new sensors and/or systems is needed;

- **the ability to report on energy consumption (EtS) of a given application**

  A simple and easy to understand metric is necessary in order to increase the user awareness of application energy consumption. EtS is one such simple metric;

- **various analysis capabilities**

  This will include calculation and reporting of application energy consumption, system power consumption, etc.

3.1. PowerDAM workflow

Fig. 3 shows the workflow of PowerDAM which is as follows:

1. The EtS is calculated for all executed applications (presented in Section 4). After the EtS calculation is done, PowerDAM updates the corresponding fields for all executed applications in the database;
2. The actual sensor data from all monitored systems is obtained;
3. Returned sensor data is checked against the existing sensors. PowerDAM logs a warning message for missing sensor data and data for unknown sensors;
4. The database is updated with actual sensor measurements;
5. Application relevant data (e.g. information on utilized compute nodes, starting/ending times of certain applications, and application owner) is obtained from the resource managers (job schedulers) of monitored HPC systems;
6. The database is updated with actual scheduler data. As soon as the database is updated, the routine is repeated (after a configurable amount of waiting time).

3.2. PowerDAM framework

PowerDAM is written in Python and provides a plug-in framework to define the monitored entities (e.g. IT systems, building...
infrastructure, etc). This makes it relatively easy to extend PowerDAM’s monitoring capabilities. PowerDAM differentiates between three functionalities: sensor data collection, scheduler (resource manager) data collection, and data analysis. It provides two plug-in interfaces for each monitored entity: one for sensor data collection and one for fetching application relevant data from resource management tools. The implementations of the interfaces define how the agent based data collection is managed.

For agent based communication PowerDAM has well-defined application programming interfaces (APIs). For example, the API syntax for sensor data collection is outlined in (1).

\[
\text{RootResourceIdentiﬁer}.\text{ResourceIdentiﬁer}^{\ast}.\text{SensorType} = \text{Value}; \text{Timestamp} \quad (1)
\]

where

- \text{RootResourceIdentiﬁer} represents the monitored entity
- \text{ResourceIdentiﬁer} represents the monitored component of the entity (e.g. in case of HPC system as a monitored entity, both a rack and a compute node can be considered as a Resource). The \(^{\ast}\) indicates that the expression enclosed in parentheses (i.e. \text{ResourceIdentiﬁer}) can be repeated (appended to the former one) zero or more times
- \text{SensorType} represents the type of the Resource sensor (e.g. power, load, temperature, etc.). Load (or utilization rate) illustrates the averaged workload of physical processors residing in the compute node
- Value represents the actual sensor measurement

![PowerDAM workflow](image1)

**Fig. 3.** PowerDAM workflow.

![Hierarchical tree structure](image2)

**Fig. 4.** Hierarchical tree structure.
Timestamp is the time and date at which the sensor measurement was obtained.

_,_,_ and ; are the delimiters between the above listed identifiers and are part of the syntax. Therefore, they can’t be used in any of the identifiers (e.g. example_example is a not valid identifier).

PowerDAM is capable of monitoring any entity which can be represented in a hierarchical tree structure due to this API syntax. Fig. 4 shows such a hierarchical structure, where each vertex represents an entity/resource and the edge indicates the relation between parent and child entities/resources.

As can be observed, each resource can have a set of different sensors. The PowerDAM API allows to easily and correctly describe the set of sensors (and their respective values at a given time stamp) which correspond to a given entity/resource and to preserve their hierarchical relationship. For example, the statement

\[
\text{SystemA.computeNodeA.power} = 232; 2013-05-24 15:34:23
\]

would indicate that the power sensor of the computeNodeA of SystemA has a value of 232 W at the time stamp value of 2013-05-24 15:34:23.

3.3. Data collection at LRZ

PowerDAM is currently deployed at LRZ and is monitoring two different HPC systems with different hardware architectures.

The first is a direct warm water cooled AMD processor based Linux cluster built by MEGWARE and called CoolMUC also referred to as Massively Parallel Processing (MPP) Linux cluster system. It has 178 nodes (2 × 8-core AMD CPU and 16 GB RAM per node) and is connected to a SorTech adsorption chiller allowing the exploration of further possibilities of waste heat reuse of the system (Fig. 5). CoolMUC uses Simple Linux Utility for Resource Management (SLURM) (Simple Linux Utility for Resource Management, 2011) as a resource management system. CoolMUC allows power monitoring for nodes, internal network equipment, and internal cooling hardware. CoolMUC has closed racks and is, therefore, room neutral (no need for room air conditioning). All heat is removed solely via the water cooling loop of LRZ.

The second system is SuperMIG (Fig. 6), the migration system for the SuperMUC supercomputer. SuperMUC was the fastest supercomputer in Europe (according to Top500 June 2012 rankings (Top500, 2013)), with 155,656 processor cores in 9288 compute nodes and a peak performance of 3 PetaFLOPS (= \(3 \times 10^{15}\) FLOPS). It uses a new form of warm water cooling technology, developed by IBM, which makes the system more energy efficient. SuperMUC consists of 18 Thin Node Islands and uses IBM LoadLeveler (Kannan et al., 2001) as a resource management system. These Thin Node Islands are equipped with Sandy Bridge-EP processors whereas the SuperMIG migration system is equipped with Westmere-EX processors (Leibniz Supercomputing Centre, 2013).

SuperMIG consists of 1 Fat Node Island with 205 nodes. Each node is equipped with 4 \(\times\) 10-core processors. This makes 40 cores per node and gives the system maximum core count of 8200. The peak performance is 78 TFlops (\(= 78 \times 10^{12}\) FLOPS).

---

Please cite this article in press as: Shoukourian, H., et al., Monitoring Power Data: A first step towards a unified energy efficiency evaluation toolset for HPC data centers, Environmental Modelling & Software (2013), http://dx.doi.org/10.1016/j.envsoft.2013.11.011
In the current setup, PowerDAM uses remote scripts to fetch the required input data from both sensors and resource management tools. Sensor data, like measurements from rack-based power distribution units (PDUs), are obtained using the system monitoring tools. Resource management specific data is obtained via SLURM (for CoolMUC) and LoadLeveler (for SuperMIG) specific commands. These scripts are integrated into PowerDAM through its plug-in framework. The collected data is put into a database for further analysis and statistics.

4. Energy-to-Solution

An important metric for PowerDAM is EtS (Auweter et al., 2011; Minartz et al., 2011). This value indicates the aggregated energy consumption of compute nodes used by the application/job and partial sub-system components (system networking, system cooling, and infrastructure) which were utilized during the run of that application. The infrastructure can be either system specific or encompass a value for multiple systems.

4.1. EtS formula

The EtS for a given finished job $J$ on system $S$ (3) is calculated iteratively as:

$$EtS(J, S) = \sum_{i = \text{startIteration}}^{\text{endIteration}} \Delta t_i \cdot P_i(J, S)$$

(3)

where

- $i$ is the iteration index
- $\text{startIteration} = \min\{j | j \cdot \text{startTime} \leq \text{timestamp}_i \leq j \cdot \text{endTime}\}
- \text{endIteration} = \max\{j | j \cdot \text{startTime} \leq \text{timestamp}_i \leq j \cdot \text{endTime}\}$

Since system cooling and networking power can only be measured for the entire system, a way has to be found to attribute a fraction of total power of such sub-systems to individual nodes (4).

The required cooling for a component can be seen as directly related to its consumed power. Therefore, a job's share of system cooling power is directly related to the proportion of its consumed power and the overall power consumed by all nodes at a given timestamp.

For calculating a job's share of the overall networking power consumption we assume that the moment the node is active its networking port is active as well. For the power consumption of the current InfiniBand network fabric the communication pattern seems to have no significant impact. This behavior was observed on SuperMUC supercomputer deployed at LRZ (Leibniz Supercomputing Centre, 2013), where the power consumption of the InfiniBand switches was relatively constant under different communication patterns. Thus, a job's networking share is a fraction of the overall system networking consumption defined by the ratio of utilized nodes and the overall number of active nodes in the system.

$$P_i(J, S) = P_i^f + \frac{P_i^s \cdot P_i^\text{cooling}}{N_i^s} + \frac{P_i^\text{networking} \cdot N_i^f}{N_i^s}$$

(4)

where

- $P_i^f$ is the power sum of all nodes which were utilized by job $J$ at the $i^{th}$ iteration of monitoring

Please cite this article in press as: Shoukourian, H., et al., Monitoring Power Data: A first step towards a unified energy efficiency evaluation toolset for HPC data centers, Environmental Modelling & Software (2013), http://dx.doi.org/10.1016/j.envsoft.2013.11.011
4.2. PowerDAM’s EtS reports

Fig. 7 illustrates one of PowerDAM’s EtS reporting options where the x-axis represents all the executed applications which were submitted to CoolMUC (MPP1) by a given user, and the y-axis represents the corresponding EtS values for each application.

The detailed EtS report showing the consumption percentages of computation and sub-system components of all submitted and executed applications for a given user is presented in Fig. 8. The first row in the bottom table of Fig. 8 presents the submitted and executed application ids. The second, third, and fourth rows correspond to the computation, infrastructure, cooling and networking consumption percentages respectively. The infrastructure percentage is currently 0% because the plug-in needed to collect the building infrastructure related sensor data was not implemented at the time of publication. Once this plug-in is implemented PowerDAM will be able to also report on other KPIs such as Power Usage Effectiveness (PUE), Energy Reuse Effectiveness (ERE), Water Usage Effectiveness (WUE), etc.

4.3. Hydro as an example application of EtS on CoolMUC and SuperMIG

Hydro (Lavallée et al., 2012) is a real-world scientific application benchmark. In contrast to synthetic benchmarks, which are developed to measure and to test certain characteristics (e.g. processor power, communication rate, etc.) of the target system, Hydro as an application benchmark, reveals a better measure of real-world performance. It solves hydrodynamics equations in 2D using an Eulerian representation on a single material and is the simplified version of the astrophysical code RAMSES (Teyssier, 2013). The problem size was set to use 160 tasks. 10 nodes were used on CoolMUC (each of nodes having 16 cores) and 4 nodes were used on SuperMIG (each of nodes having 6 cores).
used on SuperMIG (each of nodes having 40 cores), thus having the
same amount of tasks per system.

The consumed energy of the application/job executed on a given
system (as described in subsection 3.1) is calculated by PowerDAM
using the collected sensor values. The EtS can be queried by the
following command invocation (5):

\[
\text{ets /C0 system = SystemName /C0 user = userID \text{-job = jobID /C138 \text{trace = C138}} (5)
}\]

The latter --trace option, when specified, prints in addition to
aggregated EtS the detailed sensor trace of all the compute nodes
which were utilized by the given job and all sensor measurements
that were approximated.

Fig. 9 presents the EtS report of the Hydro application executed
on CoolMUC Linux cluster with enabled trace option.

The first part of the EtS report (part a) shows the sensor mea-
surements for all utilized components. The first two entries show
system level measurements as system networking power (pre-
sented as mpp1_networking_power) and system cooling power
(presented as mpp1_cooling_power) in the order of
timestamp, sensor name, value and unit. The subsequent entries show the
power and utilization rates of all compute nodes utilized by the
application. Finally, it shows the application’s share of system
networking and system cooling power presented as job.mpp1_
 networking_power and job.mpp1_cooling_power. This output format
is then repeated for the complete runtime of the application.

The second part of the EtS report (part b) lists all replacements
of invalid measurement data. In this case, an invalid utilization rate
measurement of 106.63% for lxa37 compute node was recorded.
This invalid source data was normalized to 100%.

The final part of the EtS report (part c) informs on the aggre-
gated energy consumption (EtS) and also on the consumption
percentages of the computation, networking, and cooling.

Hydro ran 1 h and 34 min on CoolMUC consuming around
4.08 kWh energy, whereas on SuperMIG it ran 47 min and
consumed approximately 2.16 kWh energy.

Fig. 10(a) and (b) present the accumulated power values of all
compute nodes which were utilized by the Hydro application
on CoolMUC and SuperMIG respectively. The y-axis of the figures
represents the accumulated power (in watts), where the x-axis
represents the corresponding monitoring timestamps. A relatively
constant power behavior of compute nodes can be observed on
both graphs. But the accumulated power of compute nodes on
SuperMIG (Fig. 10(b)) is higher than on CoolMUC (Fig. 10(a)).
Nonetheless, since it finished in half the time on SuperMIG (47 min
vs 94 min on CoolMUC), the overall energy consumption of the
Hydro application was less than on SuperMIG.

One interesting fact is that the monitored PDUs from which the
sensor data is obtained in the current system setup deliver power
average measurements in one minute intervals. Thus there is a high
probability that the timestamps which correspond to monitoring
steps of PowerDAM do not match with exact starting/ending
timestamps of applications/jobs. This results in a maximum error of

Fig. 9. EtS report of Hydro on CoolMUC.
5. Correlating data with PowerDAM

Visualization and reporting are a critical piece towards a consolidated and reliable overview of a data center. Without these options it is difficult to assess the effectiveness of proposed management solutions and even more difficult to validate, to verify, and to tune models and/or policies introduced for data center power control. For that purpose, the PowerDAM toolset provides various reporting options which further assist energy efficiency analysis of the entire HPC data center.

Fig. 11 shows the power, load (utilization), and temperature graphs for the job 42664. The graph in the upper left corner shows the accumulated power of all used compute nodes. The one in the upper right corner shows the normalized load rates of the compute nodes. The lower left and the lower right graphs show the averaged temperature behavior of each CPU of the compute nodes (on CoolMUC MPP Linux cluster, each compute node has 2 × 8-core AMD CPUs).

Since physical sensors are subject to intermittent failures or noise it is possible that some of the obtained measurement data will be invalid. PowerDAM uses linear interpolation in order to minimize the error introduced by the invalid measurements. It notifies users on the replacement of source data, if any. In Fig. 11 the cyan colored measurement points in the load graph (upper right corner) illustrate the approximated measurements for all compute nodes which were utilized by job 42664.

PowerDAM can provide a more detailed breakdown of these graphs by displaying them per utilized compute node. Fig. 12 illustrates this option for the load graph from Fig. 11. As can be observed, if there is an approximated measurement for at least one node during a certain timesamp, then this would be portrayed in the aggregated and normalized load graph (top graph in Fig. 12) as well.

PowerDAM also allows the users to see the correlation between power and load of the compute nodes which were utilized by a given application (Fig. 13). The top line illustrates the power curve, and the bottom line the load curve. This correlation can be used to classify the compute nodes into several zones (groups) according to the performance per watt indicator. This classification could be used by resource management systems for an energy efficient selection of resources. A possible example of an efficient selection policy can be the scheduling of computation-intensive applications to zones having higher performance/watt ratio and reservation of zones with lower performance/watt ratio for data-intensive applications. Another would be to schedule jobs on the most efficient zones first and on the least efficient last.

The classification map of the dynamic change of compute node sensor data through a color mapping is another analysis option provided by PowerDAM. This classification can represent any sensor type which is supported by the target compute nodes (power, load, CPU temperatures, etc.). Fig. 14 illustrates the example classification of load sensor data for compute nodes of CoolMUC MPP Linux cluster. For this example, three classification categories were defined with different ranges. The color green (rectangles marked with star) corresponds to the desired load rate of compute nodes and was mapped to illustrate the 96%–100% load range. The color white corresponds to the normal load rate of compute nodes and was mapped to illustrate the 0% and the 90%–95% load range. The color red (rectangles without star markers) corresponds to less favorable load rate of compute nodes and was mapped to illustrate the 1%–89% load range.

Other reporting options of PowerDAM include: system power consumption for a given time frame (e.g. day, month, and year).

6. Future work

Seeing that PowerDAM is only the first step towards a unified energy efficiency evaluation toolset, the following bullet points highlight future development directions:

- implementation of a plug-in required for the collection of the building infrastructure related sensor data;
- improvement of the database backend to ensure scalability for millions of sensors expected with next generation data centers and multi-peta and Exascale HPC systems;
- addition of a push communication service, where the request for data collection will be initiated by the monitored entities themselves, in contrast to current pull service, where the request for data collection is initiated by PowerDAM;

1 The star markers are being used only for illustrative purposes and are not part of PowerDAM compute node classification map.
additional reporting options to include the calculation of energy efficient key performance indicators (KPIs), such as Power Usage Effectiveness (PUE), Energy Reuse Effectiveness (ERE), Water Usage Effectiveness (WUE), Carbon Usage Effectiveness (CUE), Data Center Infrastructure Efficiency (DCiE), etc.;
- development of a web-based graphical user interface for easy access and better view of correlated data, and for providing KPI information to data center operators and policy makers;
- a public release is planned after a successful evaluation by PRACE 2IP WP11 (Partnership for Advanced Computing in Europe, Second Implementation Phase project, Work Package 11 “Prototyping”) partners.

7. Conclusion

This paper introduced the first steps towards a unified energy measurement and evaluation toolset for an HPC data center. The developed data collection and analysis tool, PowerDAM, can be used for the evaluation of energy consumption and energy efficiency of the entire HPC system and the complete data center.

Overall PowerDAM can be used to:

- classify applications according to power draw, runtime, temperature, and energy consumption

As was seen in Section 4, applications showing a high power draw can still have a low aggregated energy consumption rate due to shorter runtime. The collected data will allow to further explore the interdependence between power draw, runtime, and energy consumption;

- provide correlation between performance and energy consumption of certain applications

This correlation will further motivate the need of application performance tuning for a specific hardware architecture and allow onward exploration of static/dynamic power management techniques (e.g. voltage, current, and frequency scaling techniques);

- provide data to the resource management system (RMS) that it doesn’t currently have

For example, by reporting the EoS of applications and by providing information on the complete current state (power,
utilization, thermal, etc.) of the data center. The data could be used by the RMS for:

(a) power-cost optimization

This can be achieved by shifting energy-hungry low priority workload to off-peak hours. Thus scheduling only high priority workload (desirable with low energy consumption rate) during the peak hours and deferring low priority workload (having high energy consumption rate) to off-peak hours when the cost of electrical power is cheaper;

(b) performance per watt optimization

The power consumption of the compute node is dependent on its operational mode (sleep, idle, etc.). While there are computation intensive applications, there are also applications that require continuous I/O and/or wait for several communication messages required to accomplish computation. The CPU tends to go idle and thus consume less power during the mentioned waiting periods. The knowledge about the applications’ ETS will allow the resource manager to efficiently schedule the workload to different groups of compute nodes which have different operational modes and CPU frequencies;

- calculate energy efficiency relevant KPIs

These metrics will help to assess current energy efficiency of the data center and to verify the affect of energy efficiency improvements.

- track the status and classify the dynamic change of system compute nodes related sensor data (power, utilization, thermal, etc.)

This option will allow to track and control the violations of predefined power, utilization, thermal, etc. constraints in the HPC data center. This classification can be also useful in understanding the interdependency between the power, utilization, and temperature of a given (set of) compute node(s).

PowerDAM provides the functionalities that are needed to verify and improve existing and future data center power consumption and energy efficiency models. It allows for a wholistic view across multiple sources and structures and affords a better understanding of complete data center power profile.

Fig. 12. Breakdown for load graph from Fig. 11.
PowerDAM is and will be the underlying framework for all further investigations regarding the reduction of energy consumption and the improvement of data center energy efficiency at the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences.

Acknowledgments

The work presented here has been carried out within the PRACE First Implementation Phase project PRACE-1IP in the Work Package “Future Technologies” and PRACE Second Implementation Phase project PRACE-2IP in the Work Package “Prototyping” which have received funding from the European Community’s Seventh Framework Program (FP7/2007-2013) under grant agreements no. RI-261557 and RI-283493. The work was achieved using the PRACE Research Infrastructure resources at LRZ with support of the State of Bavaria, Germany.

The authors would like to thank Jeanette Wilde for her valuable comments and support.

References


Please cite this article in press as: Shoukourian, H., et al., Monitoring Power Data: A first step towards a unified energy efficiency evaluation toolset for HPC data centers, Environmental Modelling & Software (2013), http://dx.doi.org/10.1016/j.envsoft.2013.11.011