# GAMES: Green Active Management of Energy in IT Service centres

Massimo Bertoncini[1], Barbara Pernici[2], Ioan Salomie[3], and Stefan Wesner[4]

[1] Engineering Ingegneria Informatica, Italy
[2] Politecnico di Milano, Italy
[3] Technical University of Cluj-Napoca
[4] High Performance Computing Centre Stuttgart

**Abstract.** The vision of the recently started GAMES European Research project is a new generation of energy efficient IT Service Centres, designed taking into account both the characteristics of the applications running in the centre and context-aware adaptivity features that can be enabled both at the application level and within the IT and utility infrastructure. Adaptivity at the application is based on the service-oriented paradigm, which allows a dynamic composition and recomposition of services to guarantee Quality of Service levels that have been established with the users. At the infrastructure level, adaptivity is being sought with the capacity of switching on and off dynamically the systems components, based on the state of the service center. However, these two perspectives are usually considered separately, managing at different levels applications and infrastructure. In addition, while performance and cost are usually the main parameters being considered both during design and at run time, energy efficiency of the service centre is normally not an issue. However, given that the impact of service centres is becoming more and more important in the global energy consumption, and that energy resources, in particular in peak periods, are more and more constrained, an efficient use of energy in service centres has become an important goal. In the GAMES project, energy efficiency improvement goals are tackled based on exploiting adaptivity, on building a knowledge base for evaluating the impact of the applications on the service centre energy consumption, and exploiting the application characteristics for an improved use of resources.

## 1 Introduction

Over the last years, with the increasing digitalization of the business processes in many application domains, like online banking, e-commerce, digital entertainment, and e-health, the data centre industry has seen a great expansion due to increased need for computing capacity to support business growth. As a consequence, management of IT Processes, Systems and Data Centres has dramatically emerged as one of the most critical environmental challenges to be dealt with and new research directions are being taken towards an energy-efficient management of data centers. An estimation is reported in [10] that the

US servers and data centers consumed about 61 billion kilowatt-hours (kWh) in 2006 (1.5 percent of total U.S. electricity consumption). This estimated level of electricity consumption has been evaluated similar to the amount of electricity consumed by approximately 5.8 million average U.S. households.

In the last years, large IT systems and Data Centres are moving towards the adoption of a Service-based Model, in which the available computing resources are shared by several different users or companies. In such systems, the software is accessed as-a-service and computational capacity is provided on demand to many customers who share a pool of IT resources. The Software-As-A-Service model can provide significant economies of scale, affecting to some extent the energy efficiency of data centres. The service-based approach is becoming the most common way to provide services to users, compared to traditional application developments. Services and their composition, both at the providers' side (to provide new value-added services), and at the users' side (with mash-ups of services composed by the users themselves), are becoming more and more widespread in a variety of application domains. Hence, since the service-oriented approach is steadily increasing for many application domains, its impact on data and service centres will become more and more significant. A very similar model is applied to the provision of services in the High Performance Computing domain where users are allocated to these precious resources in a shared way using complex scheduling mechanisms.

The report [10] contains a forecast of doubling the energy consumption estimated in 2006 within five years, and it indicates that there is a potential of reducing such consumption with existing technologies and design strategies by 25 percent or more. However, many current research directions have shown that such improvement can be significantly increased considering a number of potential improvements in several aspects of a data center. Despite the big effort that has been put for assessing energy efficiency of IT service centres aiming at the reduction of energy costs [3], the most of these actions have been concerned with solutions in which energy efficiency leverages only on single, yet not interrelated factors, such as the identification of good practices for energy savings based on the dynamic management of servers according to workload and servers consolidation and virtualization; the development of low power techniques at IT component level; and the design of energy-effective facility environments for data centres through reuse of heat or air conditioning. The analysis of the characteristics of the software applications run in data centers are just starting to be considered, such as for instance in the EU best practices for data centres [4].

Mostly, these policies have been implemented in an isolated and fragmented way, not taking into account all the interrelations between the different decision-making layers and were unable to evaluate simultaneous trade-off between power, workload and performance and users' requirements. In particular, the applications running in the service centre are usually only analyzed based on their general characteristics, such as frequency of execution and requests for resources. The analysis of applications at the design level, however, could provide useful information to better manage the resources in the infrastructure. For instance, the structure of the application can be a basis for predicting the resources (e.g., data) that will be necessary for its execution. Such an information can in turn be useful for an internal management of storage resources. On the other hand, also information about IT resources can be used to design energy efficient applications. In fact, while there has been a focus on optimization and negotiation of Quality of Service and performances in the past [8, 7], very little attention has been paid to the issues of energy consumption and development of energy efficient services. A first proposal has been presented in [5], where energy consumption and energy efficiency have been considered in composed services at the same level of other quality of service parameters. This allows designing applications that can dynamically adjust to the IT infrastructure state in order to reach energy-efficiency goals, while keeping the agreed quality of service levels.

The vision of the GAMES (Green Active Management of Energy in IT Service centres) project (2010-2012) is for a Green, Real-Time and Energy-aware IT Service Centre. The central innovation sustaining the GAMES vision is that for the first time, to our knowledge, the energy efficiency of the IT Service Centres will be considered simultaneously at different levels, trading-off 1) user and functional requirements and Quality of Services versus energy costs at business/application level 2) performance, expressed as physical resources workload and Service Level Agreement, against energy costs at IT infrastructure level, 3) HVAC (Heating, Ventilating and Air Conditioning) and lighting versus the power required by the IT infrastructure and the business processes and application, as received by upper levels, at Facility level.

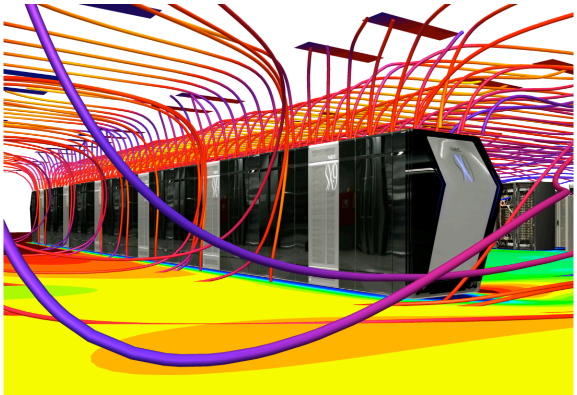

**Fig. 1.** Sample of simulation in the design phase

At design time, the assessment and benchmarking of the energy consumption and efficiency of all the different building blocks composing the GAMES IT Service Centres energy efficiency (HVAC, lighting at the facility level, servers, storage, network and processors at IT infrastructure level, services, applications, QoS) will be made for each of the sub-optimal configurations. With this regard, what-if simulation analysis will be carried out in order to determine at design time the best energy-effective distributions of services on the virtualized machines, what will be the best resource and workload configurations with less energy costs, and the impact of these configurations on the energy and carbon emissions balance of the IT Service Centre facility. Historical and required power information and the energy usage profile, combined with Business Intelligence, Data Mining and Information Extraction technologies as well as simulation technologies (e.g. Computational Fluid Dynamics simulations as shown in figure 1), will be matched with users' business, functional and applications requirements to align energy demand with availability (energy contracted with the utility operator) to design energy efficient applications on an energy efficient infrastructure, able to exploiting adaptivity during execution.

The optimized configurations, which will be the output of the GAMES system at design time, will be continuously monitored and adaptively controlled at run-time, through a suitable sensing and monitoring technology infrastructure able to measure temperature, power consumption and humidity of each single IT device (servers, storage, network). The GAMES co-design methodology will aim at co-designing business level applications and services and the IT infrastructure, to support a global energy-aware adaptive approach.

In Section 2 we illustrate the general approach to energy efficiency in GAMES, while in Sections 3 and 4 we discuss the co-design approach and the adaptive run-time environment respectively.

## 2 The GAMES approach

In the data and service centre, we envision the energy-aware design and management of service-based information systems and their IT infrastructure, supported by an adaptive SBA (Service Based Architecture), in which it is possible to dynamically modify service compositions driven by Service Level Agreements, covering Quality of Service. The goal is to realise a self-adaptive data and service centre architecture across all kind of offered resources ranging from data over computing up to the service layers. The run-time management continously balance the agreed service contracts and derive the necessary measures needed based on the monitored values (energy consumption, load situation, risk to fail on an SLA, etc.) as shown in the conceptual architecture in Figure 2.

All design choices are driven by users demands expressed as a set of Key Performance Indicators (KPI) and Green Performance Indicators (GPI) that are part of the negotiated Service Level Agreements (SLA). In order to realise this architecture, three major building blocks have been identified.
The **Energy Sensing and Monitoring Infrastructure** (ESMI) provides services to interact with the energy grid, with the environment monitoring infrastructure and with the data center resources, for energy consumption and physical measures. The ESMI has an energy service layer providing basic monitoring, messaging, event derivation features, and mining services for analysing historical data targeting the generation of useful adaptation patterns and knowledge.

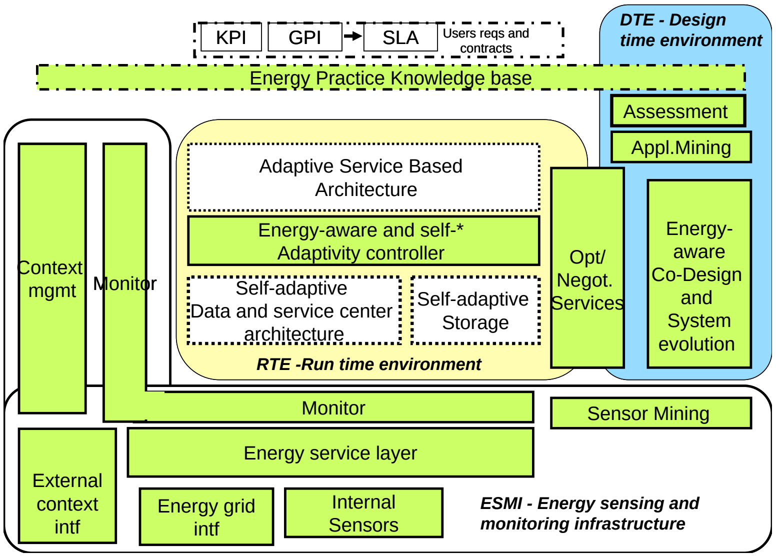
KPI
GPI
SLA
Users reqs and contracts
DTE - Design time environment
Energy Practice Knowledge base
Assessment
Appl.Mining
Adaptive Service Based Architecture
Energy-aware and self-* Adaptivity controller
Self-adaptive Data and service center architecture
Self-adaptive Storage
RTE -Run time environment
Opt/ Negot. Services
Energy-aware Co-Design and System evolution
Context mgmt
Monitor
Monitor
Sensor Mining
Energy service layer
External context intf
Energy grid intf
Internal Sensors
ESMI - Energy sensing and monitoring infrastructure

**Fig. 2.** GAMES architecture

The ESMI will be partially based on the energy service layer being developed in BeAware [6]. The sensing infrastructure will be interfaced with monitoring services, which will in addition gather relevant information from the IT infrastructure and SBA layer, generating relevant events from the sensor information. A context management support module will manage context information.

The **Run-Time Environment** (RTE) provides an energy-aware and self-* adaptivity controller including functionalities for event analysis, based on the general knowledge of the environment and energy characteristics of services, controlling the adaptivity under a global perspective of a service and an architectural level, and a general optimiser and negotiator, which, starting from static tools for architecture optimisation and SLA templates, will be enhanced with dynamic and energy-aware functionalities, exploiting also the Energy Practice Knowledge Base. The self-adaptive data centre architecture module comprises an adaptation of the architectural part and of the storage-part through strategies and decisions on data placement and storage quality of service based on access patterns and mapping of application services to data storage level.

The **Design Time Environment** (DTE) will support an energy-aware co-design of service-based information systems and IT architecture in the data and service centre. Starting from a static evaluation of existing configurations, optimisation and negotiation techniques for design time, choices will be developed, to devise the optimal functioning points to be exploited for run-time adaptivity. The design will include also the identification of the observable needs for optimal and efficient run-time event detection. Users involvement will be considered through test cases and user experience models. An assessment tool will provide an initial analysis of the users requirements, service and data characteristics and IT infrastructure and facility from which the energy-aware adaptive service and infrastructure design will start.

## 3 Designing an energy-efficient service centre

Energy-aware service-based information systems design will be tackled based on a three-fold perspective: a) strategic-level decisions in developing green IT service centres (e.g., identifying Green Key Performance Indicators (GPI) and analysing the impact of QoS business process levels on energy costs); b) control strategies to evaluate, optimise, and control services and data at run-time on multiple time scales and adapt them at run time; c) technological mechanisms and tools to reduce the energy consumption of IT service centres based on self-adaptive services and architectures. Energy savings can be obtained by exploiting the characteristics of existing adaptive platforms both at the business/application level, where adaptive service compositions can be executed, and at the architectural level, based on adaptation of IT architectures and components. The problem to be solved is how to combine the existing approaches in a layered architecture, considering a large number of information systems using the same services and sharing the same data centre(s). We propose a combined design-time and run-time approach. At design time, co-design is proposed to create adaptive service-based information systems and self-adapting architectures based on the requirements. At run-time, we propose an event-based adaptation process that takes into consideration the run-time context information (energy consumption) and design-time context information (user and business contexts).

We will focus on the design of energy-aware information systems, in which the information system functionalities and the IT system architecture are co-designed to get improved energy efficiency. The energy dimension is currently not considered in information systems design, where functionality and quality of service considerations are driving design choices. Based on some research experiments and simulation [1, 2], we advocate that considering the energy consumption dimension, different and more efficient design choices could be performed.

Examples of energy-aware co-design include not only minimized number or similar/redundant services, e.g. by using virtualisation technologies or a balanced number of servers performing supporting services operations (e.g., having only a minimal number of authentication servers) or an evaluation of the impact on needed cooling capacities based on different load scenarios of servers, but also a focus on business process analysis of core activities-services-data as shown in [9].

We will develop a cost-based approach to design the system globally and to select the adaptation strategies that are recommended at run time at the application (process/service composition) and at the IT level and to identify the variables and components which need to be monitored in order to ensure a correct control of the system. Business processes will be analyzed considering their processing requirements, data requirements and dependencies in their tasks, the ability to use alternative services in service compositions, and their context-awareness, in order to be able to enhance the adaptive capability of the application itself, but also that of the IT infrastructure, with an efficient use of the available resources as the main goal.

## 4 Energy efficiency at run time

A new approach for developing an energy-aware adaptive mechanism at run time will be defined and implemented. The basic concept is to consider and use the system context situation enhanced with energy/performance information for controlling/adjusting/enforcing the run-time energy efficiency goals. A multi-layer feedback architecture will be considered for run-time controlling of system's performance/energy ratio, by combining autonomic and context aware computing methodologies, techniques, algorithms and tools with methods and tools specific to the systems and control theory. We propose the development of different control loops that will be used to adjust and adapt the system execution to the energy efficiency goals established in the co-design phase: a set of local control loops associated to IT Infrastructure servers and one global control loop associated to the whole system. The local loop controllers are used to locally optimize the IT Infrastructure server specific energy consumption, without considering the whole system state. The local controller is developed by using a set of server specific energy optimization rules predefined at design time which can be executed on a very fine time grain without affecting the system overall performance. Using the local control loops a optimal energy consumption will be obtained for each IT Infrastructure specific component. This optimum will be communicated together with the component specific data as events to the global system controller. The global controller receives the energy-related information from each specific local loop and from the environment monitoring infrastructure as well as the performance-related information from the system's service layer in order to take adaptation decisions to enforce and realize the Key Performance Indicators (KPI) and Green Performance Indicators (GPI) defined in the co-design phase. The global control loop decision may include the execution of the following examples of energy-aware context-based adaptivity actions: minimize the necessity of calling a remote service when one local similar service is available (minimize data/service transfer), minimize the substitution of services during maintenance, optimize the number of necessary backup operations, privilege the use of services that require low energy, etc.

To derive knowledge about the service center and its energy efficiency, the GAMES framework will integrate information models that uniformly represent the system historical energy consumption related data. The general approach is based on extracting domain knowledge base from large amounts of historical data by using data mining techniques. The historical energy consumption related data will be also used together with a traceability model to understand the impact of changes in the provisioning infrastructure on energy efficiency and service quality, in order to allow both operators and consumers to select the appropriate mix as needed. With the GAMES framework it will be possible to align business requirements e.g. ”optimized for low power demand providing response time up to 200ms” versus ”optimize response time” based on historical data and the currently monitored status. By combining at design and run time the historical, predictive, context and the externally available information with the GAMES Knowledge Base will allow the selection of the most adequate adaptation patterns and profiles.

## 5 Conclusions

This paper has presented the GAMES approach to design and manage energy-efficient service centers. For implementing in a successful way the GAMES concept of energy efficiency, new overall energy efficiency metrics are needed, which will be able to assess the energy efficiency and carbon emissions in an integrated way, combining the facility with the business/process and IT architecture levels, while the most popular ones nowadays (PUE and DCiE, defined by the GreenGrid consortium [3]), are dealing only with the facility level. With this regard, the GAMES project will define and introduce new energy efficiency and emissions metrics, the GAMES Green Performance Indicators.

The general approach of co-design and adaptivity both at service and at infrastructure layer need validation, both from a theoretical point of view and from experimentation. Models and tools to be developed must be sufficiently performant and the monitoring light enough not to overload the running system. Validation in the project is planned within two large data centers, on experimental settings.

**Acknowledgments** This work has been partially supported by the European Commission within the GAMES project funded under EU FP7.

## References

1. J. Almeida, V. Almeida, D. Ardagna, C. Francalanci, and M. Trubian. Managing energy and server resources in hosting centers. In *Proc. ICAC*, 2006.
2. D. Ardagna, C. Cappiello, M. Lovera, B. Pernici, and M. Tanelli. Active energy-aware management of business-process based applications. In P. Mähönen, K. Pohl, and T. Priol, editors, *ServiceWave*, volume 5377 of *Lecture Notes in Computer Science*, pages 183–195. Springer, 2008.
3. Cristian Belady ed. Green grid data center power efficiency metrics: PUE and DCiE. 2008.
4. EU Stand-by Initiative. 2010 best practices for the eu code of conduct on data centres. Technical report, December 2009.
5. A. M. Ferreira, K. Kritikos, and B. Pernici. Energy-aware design of service-based applications. In *ICSOC 2009*, Nov. 2009.
6. L. Gamberini, L. G. Jacucci, A. Spagnolli, C. Bjorkskog, D. Kerrigan, A. Chalambalakis, L. Zamboni, G. Valentina, N. Corradi, P. Zappaterra, and G. Bosetti. Technologies to improve energy conservation in households: The users’ perspective, Maastricht. In *First European Conf. Energy Efficiency and Behaviour*, Oct. 2009.

7. P. Hasselmeyer, B. Koller, L. Schubert, and P. Wieder. Towards SLA-Supported Resource Management. In *High Performance Computing and Communications - Second International Conference, HPCC 2006, Munich, Germany, September 13-15, 2006, Proceedings*, volume 4208 of *Lecture Notes in Computer Science*, pages 743–752. Springer, 2006.
8. B. Koller and L. Schubert. Towards autonomous SLA management using a proxy-like approach. *International Journal of Multiagent and Grid Systems*, 3:313–325, 2007.
9. N.-H. Schmidt, K. Erek, L. M. Kolbe, and R. Zarnekow. Towards a Procedural Model for Sustainable Information Systems Management. In *HICSS' 09: Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10, Hawaii, USA, 2009. IEEE Computer Society.
10. U.S. Environmental Protection Agency. Report to congress on server and data center energy efficiency public law 109-431. Technical report, ENERGY STAR Program, August 2007.