VISIONARY CASE STUDIES

Green **A**ctive **M**anagement of **E**nergy in IT **S**ervice centres

GAMES

Project ICT-248514

Deliverable D7.1 WP7

Start date of project: 1 January 2010 Duration: 30 months

Report

1.0 - 30.09.2010

Document. ref.: GAMES.D7.1.HLRS.WP7.V1.0

Programme Name: .... ICT
Project Number: 248514
Project Title: .... GAMES
Partners: .... Coordinator: ENG (IT)
Contractors:
Politecnico di Milano, University of Stuttgart (HLRS), Technical University of Cluj-Napoca, IBM ISRAEL - Science and Technology LTD, Christmann informationstechnik, ENERGOECO, ENELSì.

Document Number: D7.1.V1.0
Work-Package: WP7 - Validation and Trials
Deliverable Type: Report
Contractual Date of Delivery: 30.09.2010
Actual Date of Delivery: 30.09.2010
Title of Document: Visionary Case Studies

Author(s): Jia Liu, Alexander Kipp (HLRS)
Andrea Turli, Giulio Galiero, Andrea Manieri (ENG)
Micha vor dem Berge, Wolfgang Christmann (CHRISTMANN)
Andrei Ciuca (EnergoECO)

Approval of this report: Ealan Henis (IBM), Ioan Salomie (TUC)

Summary of this report: (see Executive Summary)

History: ....

Keyword List: Visionary Case Studies, Cloud Computing, HPC Cloud, Green IT Products

Availability: public

Change History

Version	Date	Status	Author (Partner)	Description
0.1	11.08.2010	Draft	USTUTT-HLRS, ENG	Initial draft
0.2	23.08.2010	Draft	USTUTT-HLRS, POLIMI, ENG; EnergoECO	Consolidated Draft considering received comments
0.3	10.09.2010	Draft	USTUTT-HLRS	Drafted Sections 2 and 3
0.4	14.09.2010	Draft	ENG, USTUTT-HLRS, EnergoECO	Consolidated draft reflecting the received contributions and comments
0.5	15.09.2010	Draft	ENG, Christmann, USTUTT-HLRS	Consolidated Draft for Peer-Review
0.6	20.09.2010	Draft	IBM, TUC	Peer-Reviewed draft
1.0	30.09.2010	Final	ENG, Christmann, USTUTT-HLRS	Final version considering the Peer-Review

Table of Contents

Executive Summary

Within this document we're presenting visionary case studies, reflecting the challenging fields of High performance Computing (HPC), cloud computing as well as the upcoming field of provisioning of green IT products. Therefore, we first reflected the current state of the art of the corresponding fields, before describing the visionary case studies.

In general, there are many approaches and upcoming technologies, influencing the potential computing paradigms of the future, namely, quantum computing and DNA computing. However, although these approaches might have an impact on future computing paradigms the current state of these even new ways of computing are not yet mature enough to allow for a realistic analysis of the impact of these computing technologies, also taking into account the corresponding energy consumption.

Therefore, we concentrated within this document on current and actually upcoming trends in computing, namely cloud computing and HPC computing environments, whilst highlighting how these paradigms *might* emerge within the next years. Within these scenarios we also highlighted, how GAMES will allow and enable us to make these visionary case studies more realistic then they are today.

The structure of the document is as follows:

Section 1 gives a brief overview about the state of the art in current cloud computing environments, described within the frame of the well known Engineering Virtual Cloud Center (EVCC) environment, whilst highlighting how a futuristic cloud might look like.

In section 2 we describe within a brief overview the current state of the art in HPC computing environment whilst highlighting how a "futuristic", cloud enabled HPC environment might look like.

Finally, in section 3 we describe the visionary usage case about the provisioning of green IT products. This section highlights in particular how energy efficient applications and IT infrastructures can be described, monitored and steered.

1. Case Study - Cloud Computing

1.1 The current Cloud Computing approach

The current situation of cloud computing is based on a significant number of commercial cloud providers at different levels such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS)[11], offering successfully their service portfolio for quite some time. On the other hand, the industrial world, including both infrastructure providers and software developers and vendors, is being organised around virtualisation and cloud services as a private cloud. With these technologies companies can connect every small and dispersed data centre into a single virtual and federated datacentre that can be distributed around the world. Furthermore organizations can pool computational and storage power of their IT assets into the private cloud, moving easily computing power and data independently of their original locations.

The business model of federated data centres allows small and medium companies to coordinate with others for provisioning, maintenance and evolution of IT infrastructures, enabling them to face the market with an adequate efficiency level.

Basic technologies on which almost all the actors in the cloud marketplace are converging are offering the following capabilities as benefits:

- anywhere/anytime access – the cloud promises “universal” access to high-end computing and storage resources for anyone with a network device;
- specialisation and customisation of applications – the cloud offers platforms of enormous potential for building software to address a wide range of tasks and challenges;
- collaboration among users – the cloud represents an environment in which users can develop software-based services and from which they can deliver them;
- processing power on demand – the cloud is an “always-on” computing resource that enables users to tailor consumption to their specific needs;
- storage as a universal service – the cloud represents a remote but scalable storage resource for users anywhere and everywhere;
- cost benefits – the cloud promises to deliver computing power and services at a lower cost than installing and maintaining your own hardware.

The adoption of cloud computing model contains a number of open issue. To give some examples: (i) the companies hosting the Cloud services control, and thus, can monitor the communication and data stored between the user and the host company (data privacy concern), (ii) the EU Data Protection [24] and the credit card industry's PCI DSS [25] could be difficult to audit in cloud environment (compliance concern).

1.2 The Cloud Computing reference scenarios

Engineering SpA will provide a cloud prototype solution for a Virtual Cloud Centre that is a self-contained bundle of software (SW) allowing small and medium enterprises and small data centers to easily switch to cloud computing. The first version will have a basic set of functionality that will evolve in the near future to the full model described in the following paragraphs.

The Virtual Cloud Centre allows for the full management of the registered computing and storage resources, their monitoring and accounting at different granularity and a model for federating authentication and authorization of different applications.

Engineering Virtual Cloud Center (EVCC) [12] provides an Open Cloud Computing Interface (OCCI) interfaces and can interoperate with an ecosystem of other EVCCs moving data and computing resources.

EVCC uses best-of-breed virtualization techniques, network-, compute-, storage-, security-, and management-products, guaranteeing optimal integration with standards and Open Source context.

Hosting EVCC in the Engineering Data Center offers advantages of fully covering the Information Technology Infrastructure Library (ITIL) service management processes for customers.

When a customer installs EVCC in his own data center, EVCC guarantees Service Level Agreements (SLAs), due to its predictable performance and operational characteristics, and reduces risks and compliance issues because it is a tested and validated solution with unified support and end-to-end vendor accountability.

EVCC is a suite of integrated components consisting of two basic services for computation (EVCC Computational component) and storage (EVCC Storage), a bottom layer for the virtualization technology abstraction (EVCC Virtualization software) and additional layers for cross-cutting functionalities such as: security management (EVCC Security Management), accounting management (EVCC Accounting Management) and the data center monitoring (EVCC Monitoring software).

A detailed description of each component is provided below:

- EVCC Computational component is the main building block. The architecture is based on the CX1000 cloud server infrastructure by Fujitsu [20] allowing simultaneously i) powerful scale out performance up to hundreds of nodes; ii) low footprint in space and power requirements; iii) agility in performing automatic provisioning, de-provisioning and repurposing of physical and virtual machines; iv) flexible network configuration allowing to wire once the infrastructure.
- EVCC Storage is a basic building block that: (i) offers a flexible and configurable connection to the storage, (ii) permits to configure multiple storage backend out of the box: NFS, SHH, LVM, that can easily be extended through plugins; (iii) offers server-less backup and integrated snapshotting and cloning features; (iv) can evolve with the capability of activating Disaster Recovery (DR) features.
- EVCC Monitoring software is a component that offers monitoring and management capabilities derived from Engineering's long experience in data center management and monitoring. EVCC can use these enhanced monitoring functions through the use of centralized Engineering Monitoring Tools, or can use basic monitoring fully integrated with its components.
- EVCC Virtualization software can use different Hypervisors like KVM, Xen, and VMWare. Cloud Management software offers a high level set of functionalities making EVCC easy to enhance, to support new functionality, and to embed into other cloud applications and platforms. It offers a uniform view of the resource pool, a centralized orchestration, a centralized network management, and a fully comprehensive Virtual Machine life cycle management.
- EVCC Security management has a multi-tenant administration, role-based security, and strong user authentication, federated security Authorization and Authentication.
- EVCC Accounting management is a building block that allows to register all major events in the Cloud infrastructure, and to measure resource consumption in order

to provide billing information for customers.

The Cloud Computing case study will be experimented in an EVCC-based testbed environment where two different scenarios will be evaluated: (1) a data-intensive and (2) a computing-intensive processing. The former will deal with the deployment and execution of a broad set of services developed for D4Science, an EU project related to distributed digital libraries [13]. Such services have been specifically implemented in order to deal with very large data-sets. The latter will deal with the deployment and execution of a set of services developed for ETICS, an EU project related to distributed software building and testing[14]. The processing of ETICS services is extremely CPU-bound, since a great computational power is needed in order to compile the sources, package the binaries and execute the tests against the artefacts created for hundreds/thousands of software components on a distributed platform. Thus, both the D4Science and the ETICS scenarios will be deployed on top of the EVCC as our cloud testing environment. With reference to the GAMES project, valuable feedback will be gathered from the analysis of data regarding the energy consumption of a cloud-based data centre in the two extremely different scenarios.

1.3 The futuristic Cloud Computing approach

There are many limitations in the current cloud scenario that will be a challenge in the near future. Cloud computing is facing the evolution of virtualization technology, in terms of using a unique e-infrastructure that guarantees independency, optimization of resource utilization, reliability, energy efficiency and maintenance costs. The development of a cloud security platform that totally guarantees isolation and safety of user data at all functional layers is an important challenge faced by many cloud technology providers.

To maximize the benefits of cloud computing and to increase interoperability with multiple clouds at different functional layers, it is crucial to adopt common open standards, following leading standardization bodies such as e.g. Open Grid Forum (OGF[15]), SNIA, SNIA-Europe[16] and the Distributed Management Task Force (DMTF[17]). In addition, the interoperability among open source and commercial solutions, will allow users to flexibly move workloads across different clouds.

Cloud platforms should offer cost models that are flexible and easily adaptable to various needs. IT Service level management in the cloud needs to adhere to standards and methods as described in ITIL and in other common international reference models (ISO20000 [22], COBIT [21]), in order to automatically integrate service management requests and measures into cloud operations.

Along with these challenges, a futuristic cloud computing approach should also demonstrate to follow a green computing approach, by offering concrete metrics both on energy consumption and environmental effects.

<u>IaaS</u>

- federated data centre for heterogeneous multicore platforms to follow and support the evolution of architectures for parallel computing and programming;
- data centre services that allow interoperability between clouds offered by different providers;
- enhancement of technologies and procedures that enable network automation provisioning, simple VM provisioning or more complex network layouts that a customer may require for a set of VMs;

- identification of all HW and SW techniques, as well as strategies and metrics to improve the energy efficiency of existing data centres and of new smart data centres;
- study of models and open source techniques to federate data centres.

PaaS

- security enhancements for authorization and authentication techniques in a federated environment; security enhancement for data isolation;
- enhancements for security logging and tracing as well as monitoring for distributed clouds;
- accounting and billing in federated clouds;
- models and methods for the mash-up of services and heterogeneous data at run-time;
- models and methods to support applications with massive data to be managed and moved across distributed clouds;
- programming models and languages to be developed, tested and executed on multicore platforms, with a development environment that enables to reduce the learning curve;
- solutions for automatic provisioning of applications in the cloud as PaaS;
- compliance needs for the cloud platform for security certifications such as ISO 27001 [23], PCI DSS [25], and International rules in terms of privacy.

SaaS

- experiment models and solutions to deliver cloud based ICT services with simplified access, on mobile devices (smart phones, iPhones/iPad), focusing on some reference application for multimedia delivery, environment monitoring, invoicing, logistic etc.

2. Case Study - High Performance Computing

High performance computing plays an incomparable role in industrial areas and academic researches, particularly for compute intensive applications. Many references in the literature, e.g. [18][19], have described the general features of HPC and emphasized its profound meaning for modern daily life. For this reason, the functionality and architecture are not the focus of this section. This section will concentrate on the developing tendency of HPC, in particular in the context of GAMES.

2.1 Current business procedure of HPC

Although the requirements for HPC from industrial areas and academic areas are increasing steadily, it is conceivable that not every company or institute is able to afford to build its own HPC center as desired. The enormous investment is one of the major barriers. In addition, since a HPC center is a notorious power consumer, the high costs of electric power and maintenance are also heavy burdens. Consequently, the alternative solution for most users is to rent computing resources from available HPC centers when they are needed. As a result, they become the main clients of these HPC centers. For instance, as one of the three largest HPC centers in Germany, HLRS (High Performance Computing Center Stuttgart) owns several different high performance cluster systems, allowing the support of differing needs [1]. It is able to provide "thousands of cores" supercomputing capacity. Many industrial partners have long term cooperation relationship with HLRS. Unfortunately, the lease of computing resources is not as easy as renting a car or a house. Technical and economical factors are absolutely determinative. Hence, HLRS does not only offer the computing resources, but also provides consulting services. Nevertheless, the current business procedure would sometimes lead to inevitable problems. In this section, the main problems will be introduced.

Under the current business procedure, given that users have requirements for HPC resources, the first problem they will suffer from is how to decide on a suitable configuration of a system for their applications. Indeed, to configure the according computing resource triggers many relevant problems. Examples of questions that need to be addressed are: how many cores are actually needed, how much memory is required for computation, how should the machines be configured, and so on. Furthermore, there are also many other involved technical problems affecting the performance of HPC systems. There is no subjective norm to assess the configuration; it depends on the specific requirements of the according users. In most cases, it is difficult to customize the optimal configuration for applications. The following situations are undesired: "overestimated" configuration will occupy unnecessary computing resources. It is obviously also opposite to the Green IT concept. The unnecessary devices are inducing further waste of resources and thus consume additional electric power, although they would be mostly operating in idle mode. On the other hand, "underestimated" configuration means oppositely that the reserved computing resources are not sufficient to allow for an optimized execution of the application. This situation has even worse results, because the system has to be reconfigured due to insufficient computing resources, since otherwise, the application would be conceivably executed for unacceptable long time. As a conclusion, in order to determine a suitable configuration, solid HPC knowledge is definitely required. However, the general users do not want to be bothered with these technical problems. Therefore, the complex process is expected to be simplified and automatically done by the systems themselves. In other words, the process should be completely transparent to the users.

Besides the intractable technical problem, another issue is also unavoidable: the operation duration. Users have to predict how long they will rent clusters and other involved devices. Undoubtedly, this time frame relies strongly on the corresponding applications and the computing resources. It is difficult to predict or estimate the exact time frame for the applications running their workloads. Therefore, most users end up reserving the system for longer durations than what they actually need. As a result, the computing resources occasionally stay in low usage or in idle mode. From the management point of view of a HPC centre, predefining the time frame is easy to schedule, arrange and coordinate, among numerous reservations. On the other hand, the fixed time frame issue is also an obstacle to improve the energy efficiency in HPC centers.

In summary, it is apparent that the manual customization of the computing systems results in incredible inconvenience for users. Furthermore, the time frame for resource reservation does not only cause additional expenditure, but also limits the usage levels of computing resources. In the context of Green IT, the current business procedure needs improvement. To overcome the drawbacks, many HPC centers are starting to find alternative business procedures.

2.2 The current business model in cloud computing

The current business procedure for HPC centers and its main drawbacks have been stated in the above section. Before looking into the futuristic HPC scenario, a significant application, cloud computing, has to be introduced here briefly. The Cloud computing paradigm has become popular in the last decade. While cloud computing is developing and maturing gradually, more and more services are provided in clouds, including even high performance computing services. Meanwhile, its successful experience and paradigms also inspire HPC providers. Analogous with the last section, this section also sets the key point on the business models of cloud computing and Amazon's service Elastic Compute Cloud (EC2) [2] is taken as an example.

As one of the main cloud services, EC2 is aiming to provide resizable virtual computing capacity in the cloud by widely adopting virtual machine techniques [2]. Hence, the mentioned configuration in EC2 is in fact a virtual configuration. Each virtual machine, called an "instance", functions as a virtual private server. The service enables the users to apply web service interfaces to launch instances with a variety of operating systems, load them with a customized application environment, manage the network's access permissions, and run the image using as many systems as they desire[2]. Table 1 summarizes the virtual configuration categories of Amazon EC2. Moreover, each category is refined into multiple options to fulfill different applications/users' requirements. Evidently, each resource group has its own price, depending on the offered configuration type.

Users are enabled to choose a suitable configuration category with the consideration of the characteristics of applications, their own requirements and also the price factors, but they do not need to be concerned about the physical clusters or storages. Their applications are easily replicated from one cluster to another. The cloud providers can effectively improve the usages of clusters. Due to the property of virtual machines, the tasks and processes can be immigrated to some clusters, and other nodes / clusters that are in low usage or idle can be turned to sleep, or even shut down. Some users do need their computing resource just temporarily, and the according virtual machines can then be stopped. Hence, the physical computing resources can be re-deployed for other VMs. In this way, the providers and users can benefit from each other reciprocally.

During the entire process, the physical computing resources are completely transparent for the users. The operation duration is also dynamic. Both main drawbacks of the current HPC procedure do not exist in this cloud scenario.

Standard resource set	This group is well suited for most applications, the optional configurations are as follows:
	Small set as default configuration, 1.7 GB memory, 1 EC2 compute unit (1 virtual core with 1 EC2 compute unit), 160 GB local storage, 32 bit platform
	Large set: 7.5 GB memory, 4 EC2 compute units (2 virtual cores with 2 EC2 compute units each), 850 GB of local storage , 64 bit platform
	Extra large set: 15GB memory, 8 EC2 compute units (4 virtual cores with 2 EC2 compute units each), 1690 GB local storage, 64 bit platform
High-memory set	Large memory size is offered in this set for high throughput applications, including database and memory caching applications:
	High memory extra large set: 17.1 GB memory, 6.5 EC2 compute units (2 virtual cores with 3.25 EC2 compute units each), 420 GB local storage, 64 bit platform
	High memory double extra large set: 34.2 GB memory, 13 EC2 compute units (4 virtual cores with 3.25 EC2 compute unit each), 850 GB local storage, 64 bit platform
	High memory quadruple extra large set: 68.4 GB memory, 26 EC2 compute units (8 virtual cores with 3.25 EC2 compute units each), 1690 GB local storage, 64 bit platform
High-CPU set	The members in this set possess proportionally more computing resources than memory and are well suited for compute intensive applications:
	High CPU medium set: 1.7 GB memory, 5 EC2 compute units (2 virtual cores with 2.5 EC2 compute Unites each), 350 GB local storage, 32 bit platform
	High CPU extra large set: 7 GB memory, 20 EC2 compute units (8 virtual cores with 2.5 EC2 compute units each), 1690 GB local storage, 64 bit platform
Cluster compute	This set provides proportionally high CPU with increased network

set	performance and are well suited for high performance computing applications and other demanding network bound applications:
	Cluster compute quadruple extra large: 23 GB memory, 33.5 EC2 compute units, 1690 GB local instance storage, 64 bit platform, 10 Gigabit Ethernet

Table 1 - EC2 Instance types (Source: [2])

Among the mentioned configuration groups, the cluster compute set arouses our interest. This category is aiming at HPC applications; in particular, the cloud providers offer specific designed clusters and a high performance network in order to meet the requirements of HPC applications. This is particularly valuable for those applications that rely on special protocols like Message Passing Interface for tightly coupled inter-node communication, which enforce some special requirements to the according IT infrastructure, like increased bandwidth and reliability [2]. Meanwhile, a controversy, whether Amazon's cluster computer service can replace real HPC center in future, is becoming fierce. At present, it is difficult to draw a conclusion. Although cloud computing provides an extensible and powerful computing environment for web services[3], the techniques and paradigms of the cloud HPC service are not yet mature enough for typical HPC computations. Two key factors constrain the development of the cloud HPC service. First, the current computing resources are not sufficient for most HPC applications. According to the configuration list, the offered computing system is still far from the requirement of general HPC applications. In addition, although cloud providers offering HPC services guarantee high performance network connections for this service, the performance of this network is hard to compare with Infiniband or Cray SeaStar interconnections, which are widely adopted in general HPC centers. Unless these two barriers are hurdled, the cloud HPC service will not be able to replace real HPC centers.

2.3 The futuristic HPC business model (cloud HPC)

The business model of cloud computing has been introduced in the last section. Its advantages have also been emphasized. It cannot be denied that both users and service providers benefit from the model. For users, it simplifies the complicated process of determining the suitable configurations for their applications. They also only need to pay for the time period that the according resources have been used. In addition, providers can effectively get rid of their low usage clusters. Consequently, the capacity utilization and hereby the efficiency can be dramatically enhanced, which leads to reduced overall energy consumption.

In principle, the business model of cloud computing can be transplanted to HPC. HPC owners would also like to accept this "new model". Therefore, a new concept "HPC Cloud" [5] has been recently proposed. This concept simplifies the determination of the system configuration on the user side. In the near future, the business model of HPC is expected to provide some predefined configuration categories for general purpose applications. However, it does not completely remove the option to customize computing systems by users. Users are enabled to choose systems more flexibly from the predefined categories with the fully hidden infrastructure, or can customize the systems with fully open resources, in the same manner as with the classic HPC. Additionally, for each category, users can be informed about the predicted time period required to run their workload, and about the corresponding energy cost. This information may be displayed visually and be based on the Green Performance Indicators (GPIs). In the context of Green IT, the green parameters, e.g. the energy consumption, will become a high priority attribute during system configuration. Instead of contracts over a fixed period of time, users only need to pay for what they used, in a so called “pay per use” manner. Hence, users are enabled to completely control the resource consumption according to their budget, which is a main economic contribution of the new business model.

Besides the benefits on the user side, the HPC owners have a chance to reduce the energy consumption of HPC environments. In the context of GAMES, the energy efficiency, described, e.g., by Green Performance Indicators (GPIs), could play a more important role than before, and can be used as a crucial factor to evaluate the system performance, besides the Key Performance Indicators (KPIs). Therefore, HPC owners can classify computing capacity into several configuration categories based on the energy consumption, (examples are presented in Table below), similar to the categories of cloud computing shown above in Table 1. Similar to Table 1, the categories in Table 2 are targeted on different potential requirements of HPC applications. However, the main difference between the 2 tables is that the configurations mentioned in Table 1 refer to the configurations of virtual machines. In contrast, Table 2 presents the according, concrete hardware configurations. In addition, it should be noted that each category can also be further refined into several subgroups to meet different requirements. In this section we concentrated on giving an overview.

Standard group	Similar to the standard resource set in Table 1, this group is also oriented to satisfy the requirements of most general applications. However, the configurations in the corresponding sort in Table 1 are not sufficient for HPC applications.
High Performance Computing group	This group offers the configuration with high computing capacity. The performance can be guaranteed. The energy consumption is high. The anticipated time period is short. The cost is high. It is suitable for users with special requirements for the operation period. Analogous to High CPU set, the High Performance Computing group is also targeting to fulfill the relevant requirements regarding computing resources of HPC applications.
High Memory group	Large memory space is provided, same to High-Memory set in Table 1. It is oriented to users with specific requirements for memory or storage systems.
High I/O group	This group is supposed to serve the applications with high data communication requirements. High speed networks are adopted .

Table 2 - The classification of energy profile of HPC centers

It seems that the “HPC cloud” potentially combines the HPC center and cloud computing approaches successfully. Unfortunately, some obstacles cannot be simply dodged. These problems are caused by the difference between their orientations and features.

- *Elasticity* is the most important property of cloud computing and circumscribes the capability of the underlying infrastructure to adapt to changing, potentially non functional requirements. For example, the amount and size of data supported by an application, the number of concurrent users etc. may change [4]. However, elasticity is difficult to adopt in HPC environments. Current HPC environments do not allow for the dynamic integration and extraction of physical resources to the infrastructure for the replication of services / instances.

- *Reliability*. As the essential feature of the cloud systems, reliability denotes the capability to ensure constant operation of the system without disruption, i.e., no loss of data, no code resets during execution etc [4][5].
- *Availability* enforces to introduce redundancy for services and data so failures can be masked transparently. Cloud providers have to ensure that their services and data are accessible from anywhere at any time [4]. In contrast, the availability of HPC resources on demand can only be granted as long as there is no broad uptake of the model. In other words, this is only possible as long as the demand does not exceed available resources [5]. In fact, the seemingly infinite resources of the cloud are an illusion as seen by the user. The cloud resources have their (although large relative to a typical user's requirements) finite limits. Hence, the difference from HPC in this respect is quantitative and not qualitative.

In the following concrete use case, the stated properties of the futuristic HPC business model will be explained. As an example, let us assume that a user intends to submit a job into the HPC cloud. The HPC cloud system provides 4 possible options, as presented in Table 2, per an interface. Assuming that the job is a computing intensive job, and that the user has a strict time requirements. For these reasons, the High Performance Computing group has been selected. Sequentially, the system reports to the user the estimated time period, energy cost and possible expenditure, inferred from analyzing the complexity of the submitted job and other related parameters. After the confirmation of the user, the system starts to execute the job. During the execution, the user can monitor the online parameters, e.g., the remaining time, the current energy consumption and so on. Meanwhile, the system logs its status periodically. It is also expected that the user is able to change the system configuration and categories online, in order to adapt to changes of the user's requirements. After redefining, all mentioned parameters, the predicted energy consumption, time period, and other parameters of interest will be re-estimated and sent as feedback to the user.

So far, we have given an overview of the futuristic HPC business procedure and a concrete example. The futuristic procedure takes full advantage of the cloud computing features to hide the complex system configurations and additional relevant issues from the users, as stated. At the same time, it also provides a user friendly way to access, use and monitor the according resources. Therefore, users are encouraged and able to easily use the computing resources of HPC centers. Perceivably, with the simplified procedure, users only need to submit their applications and decide on the energy profile as presented in Table , in the context of Green IT. All other awkward issues, e.g., assigning the jobs to clusters/computing units, the types of operation systems etc. are completely transparent for the users. In the background, computing resources will be automatically allocated and configured by the management procedures according to the selected energy profile and the power requirements of the various devices.

In the discussion above, the main advantages and barriers, which the new concept brings, have been elaborated. The main source for the hurdles is the difference in orientation of cloud computing and HPC. But still, the proposed "HPC cloud" concept illustrates a good perspective for integrating HPC and cloud technologies.

3. Case Study - Provision of Green IT Products

As one of the hottest topics in current IT industry, Green IT has aroused the interest of governments, researchers and hardware vendors. Many researches and projects addressing IT energy consumption have been conducted. Consequently, new techniques and methods are also developed. Among them, many efforts have been investigated in order to developed hardware to curb the increasing power requirement. For example, Dynamic Frequency Scaling [20] technique is one of the most famous, whereby the frequency of a microprocessor can be automatically adjusted "on-the-fly" according to the usage level of the CPU, both to conserve power and to reduce the amount of heat generated by the chip. Moreover, the new techniques of storage systems, monitors and other parts are also utilized to save energy. Besides the technique development, Green frameworks are also a potential measure for Green IT. In this chapter, we are focussing on futuristic Green IT products with Green IT frameworks, in particular with GAMES.

Before presenting the perspective of futuristic Green products, let us revisit the project GAMES, in particular the GPIs and the monitoring techniques. GPIs (Green Performance Indicators)[26] encompass many different metrics, spreading from low level (e.g., infrastructure) to high level (including middleware layer and application layer). In fact, GPIs pave the way to subjectively assess a crucial issue: how green systems are. Nonetheless, assessing a system by employing GPIs, the data that represent the instantaneous system states become implicitly necessary. In GAMES, the data is collected by monitoring techniques before being analysed and provided to the relevant modules in order to transform the GPIs instantaneously in different levels. The according GPIs are called instant GPIs. Evidently, it is unfeasible to measure all the parameters in the according systems. Selecting the dominant parameters from them provides a practical way, as listed in [27]. Hitherto, the background knowledge of GPIs and monitoring techniques has been reviewed briefly.

At present, people are used to determining the performance level which they want/need, when they are using general home appliances, e.g., refrigerators, washing machines and others. As the most authoritative criterion, the European Union introduced an energy label, widely accepted by many manufacturers [10], allowing the users to determine the energy efficiency of the according products.

Actually, the home appliances offer some degree of flexibility to users; they also provide a good model for future Green IT products. In the futuristic Green IT model, it is expected that users are not only able to choose the performance level, but also to be informed about the according energy consumption and efficiency. Furthermore, users are allowed to decide the green performance degrees, similar to deciding the performance levels.

Let us take an example to elaborate. A service provider / hardware vendor would supply a service/product with some possible green performance grades for most general requirements, e.g., from high performance computing with costly high energy consumption to low energy consumption, but with lower performance. Between the two extremes, several grades are predefined with regard to general requirements. As stated, the green performance can be represented by GPIs. Therefore, for each green performance degree, the according set of GPIs would be defined as the norms or reference values. It can be perceivable that occasionally the energy consumption of services / products does not fulfil the chosen green performance degree. It can be also interpreted that the instant GPIs do not satisfy the reference GPIs. Due to the GPI difference between real value and the reference value, the adaptive controller will be triggered to determine the adaptation actions that need to be executed to enforce the GPIs. Meanwhile, the monitoring plug-in/sensors continue collecting the immediate information of the systems and send back to gauge the GPIs. When the instant GPIs meet the requirements of the predefined GPIs, the controller would be going to “Stand By” mode again. Of course, the entire control process, including information collection, the comparison, controller activation/deactivation, is conducted in the background and is transparent to users. Besides the above mentioned predefined green performance degrees, the providers also allow users to customise the configuration of systems due to some specific requirements, similar to current products. However, in addition information about the energy consumption of specific systems will also be made available to users.

On the users’ side, the system selecting process has been simplified and the energy consumption has become visible. The only thing users have to do is to determine the energy consumption grade according to their specific requirements. As an example, when a user only pursues high performance regardless of any energy consumption consideration, the high performance grade is the first choice. On the contrary, when energy consumption is the high priority factor and the performance is not essential, the high energy efficiency should be chosen. It has to be noted that the performance and the energy consumption are indeed proportional. As mentioned, it means that the high energy consumption must be the price for the high performance, and vice versa. Consequentially, it is impossible to require high performance and also low energy consumption at the same time. Therefore, users have to balance the energy/performance trade-off according to their requirements.

Besides the Green framework contributions, green certificates are also a necessary issue for Green IT. Currently, Energy Star and Electronic Product Environmental Assessment Tool (EPEAT) are the most authoritative certificates to label the qualified products. In the following, these two certificates will be introduced concisely.

Energy Star is the trusted, government backed symbol for energy efficiency, helping to save money and protect the environment through energy efficient products and practices. The Energy Star label is targeting to reduce greenhouse gas emissions and other pollutants caused by inefficient use of energy, and brings convenience to users by identifying energy efficient products that offer savings on energy bills without sacrificing performance, features and comfort [7][7]. Based on the specific features of each product, Energy Star established different requirements. Taking the computer category as an example, this category has been further refined into several subsections: desktops & Integrated Computers, Notebooks and Tablets, Workstations, Small Scale Servers and Thin Clients. It drafts the following four criteria to assess the energy efficiency:

- Power Supply and Power management criteria
- Efficiency and performance requirements
- Power management requirements
- Voluntary requirements

As stated, Energy Star uses the four criteria also to evaluate the energy efficiency of the according computer category. Further details of the criteria are presented in [9]. According to statistic results [7][7], an energy star qualified computer meeting the Energy Star specification consumes between 30% and 65% less energy, depending on how it is used.

Productions can earn the Energy Star label by meeting the energy efficiency requirements set forth in the Energy Star production specifications. EPA (United States Environmental Protection Agency) establishes specifications based on the following set of key guiding principles [7][7]:

- Product categories must contribute significant energy saving nationwide.
- Qualified products must deliver the features and performance demanded by

consumers in addition to increased energy efficiency.

- If the qualified production costs more than a conventional, less efficient counterpart, purchasers will recover their investment in increased energy efficiency through utility bill saving, within a reasonable period of time.
- Energy efficiency can be achieved through broadly available, non-proprietary technologies offered by more than one manufacturer.
- Product energy consumption and performance can be measured and verified with testing
- Labelling would effectively differentiate products and be visible for purchases.

EPEAT[8], another famous energy efficiency certificate, is a system that helps purchasers evaluate, compare and select electronic products based on their environmental attributes. Comparing with Energy Star, EPEAT does not cover many products, rather it is specific for evaluating desktop and laptop computers, thin clients, workstations and computer monitors. EPEAT's consists of 23 energy oriented required criteria and 28 optional criteria, which are sorted into 8 groups as follows:

- Reduction/elimination of environmentally sensitive materials
- Materials selection
- Design for end of life
- Product longevity/life cycle extension
- Energy conservation
- End of life management
- Corporation performance
- Packaging

Furthermore, EPEAT employs the following rating system to evaluate the energy performance.

- Bronze- product meets all 23 required criteria
- Silver- product meets all 23 required criteria plus at least 50% the optional criteria
- Gold – production meet all 23 required criteria plus at least 75% the optional criteria

It has to be noted that the criterion "Energy conservation" is in fact based on the Energy Star classification. Apparently, EPEAT does not focus on the evaluation of energy consumption during operating a product, but rather on the entire lifecycle, including manufacturing, packaging and disposal.

According to the green electronics council's EPEAT 2007 Environmental Benefits report, the following savings were realized in 2007 [6][6]:

- 42.2 billion kWh of electricity were saved
- 174 million metric ton (including 3.31 million metric tons of greenhouse gas) were eliminated
- 365,000 metric tons of water pollutant emissions were eliminated
- Financial savings for companies will total more than 3.6 billion US dollars over the life of EPEAT products sold in 2007, primarily from reduced energy usage.

Clearly, energy consumption can be dramatically reduced through exploiting energy certificates. Unfortunately, none of the above mentioned two certificates are really oriented to specifically assess the IT products and IT services. Therefore, it is necessary to draft a new energy efficiency certificate. We propose a new Green IT certificate, which should be based on the successful experience of Energy Star and EPEAT. However, it should not be limited to assessing the energy consumption, but will also be extended to take GPIs into consideration. Similar to the EPEAT certificate, the Green IT certificate will also employ a rating system to group the energy efficiency situation of IT services into different categories, e.g., from highest efficiency to lowest efficiency. It will also be possible to refine it further. For popularizing the new certificate, the governments have to devote some efforts. Of course, the certificate will allow users easier realization of how much energy their selected services would consume, without any complex descriptions, and in a visible and understandable way.

With the Green IT certificate and the corresponding business model, users will be enabled to easily select an appropriate grade. Conceivably, when users submit their applications/jobs, they should be notified of the energy consumption of each grade to execute their applications, associated with high performance or low energy cost. Thus, users will be free to choose the grade which is most suitable for their requirements, taking time and cost into consideration. In addition, the business model, together with the Green IT certificate, is able to balance the time/cost trade-off in a reasonable way.

4. Conclusions

Within this document we have presented 3 challenging visionary case studies about potential futuristic computing environments, whilst highlighting the energy consumption and efficiency of the according environments.

It has been suggested and argued that the GAMES approach can be an enabler of the described visionary case studies, in particular allowing for the description of applications and IT infrastructures properties related to their energy consumption and energy efficiency. The GAMES approach will hopefully allow to make a significant step towards this visionary usage case, in particular by:

- providing the definition of GPIs and KPIs relating to the according energy consumption and efficiency at different levels of applications and IT infrastructure,
- linking and relating these GPIs and KPIs via different levels,
- providing a design-time environment allowing for the design of applications and IT infrastructures whilst taking into account the according energy consumption and efficiency
- offering adaptive run-time controllers that determine the best adaptation actions to be taken to enforce the energy efficiency goals.
- allowing for the description of energy aware applications and IT infrastructures in a transparent and traceable way

5. References

[1] HLRS website. Available at http://www.hlrs.de/
[2] EC2 introduction from Amazon. Available at http://aws.amazon.com/ec2/
[3] Jeffrey Napper, Paolo Bientinesi. Can Cloud Computing Reach The Top 500
[4] Lutz Schubert. The future of Cloud Computing: Opportunities for European Cloud Computing Beyond 2010
[5] Lutz Schubert, Stefan Wesner. HPC Clouds: Cloud Computing from the perspective of a HPC provider. Zakopane, Poland, March 2010
[6] Green Electronics Council, Environmental benefits of 2007 EPEAT purchasing: Green IT procurement System's Success Drives Major Environmental Benefits. Available at http://www.epeat.net/Docs/EnvironmentalBenefits2007.pdf
[7] The official website of Energy Star. Available at http://www.energystar.gov
[8] The official website of EPEAT. Available at http://www.epeat.net/default.aspx
[9] Energy Star Program Requirements for Computers version 5.0. Available at http://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer/Version5.0_Computer_Spec.pdf
[10] European Union energy label. Available at http://en.wikipedia.org/wiki/European_Union_energy_label
[11] Youseff L, Butrico M, Da Silva D. *Toward a Unified Ontology of Cloud Computing*. IEEE; 2008:1-10. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4738443.
[12] Engineering Virtual Cloud Center (EVCC): Internal Documentation
[13] The D4Science project: http://www.d4science.eu
[14] The ETICS project: http://etics.web.cern.ch/etics/
[15] Open Grid Forum www.ogf.org
[16] The Storage Networking Industry Association, www.snia-europe.org
[17] Distributed Management Task Force - www.dmtf.org
[18] Yang LT, Guo M. *High performance computing: paradigm and infrastructure*. Wiley-Interscience; 2005:816.
[19] Gruber R, Keller V. *HPC@Green IT: Green High performance computing methods*. 1 ed. Berlin: Springer; 2010:215.
[20] Fujitsu CX1000 primergy: http://ts.fujitsu.com/products/standard_servers/scale-out_servers/cx1000/primergy_cx1000.html
[21] COBIT: http://en.wikipedia.org/wiki/COBIT
[22] ISO 2000: http://en.wikipedia.org/wiki/ISO/IEC_20000
[23] ISO27001: http://en.wikipedia.org/wiki/ISO/IEC_27001
[24] Data Protection Directive: http://en.wikipedia.org/wiki/Data_Protection_Directive
[25] PCI: https://www.pcisecuritystandards.org/security_standards/pci_dss.shtml
[26] GAMES Deliverable D2.1 - Layered Green Performance Indicators Definition
[27] GAMES Deliverable D5.2 - First Stage Sensor Data Analysis