



Business-driven management of infrastructure-level risks in Cloud providers

J. Oriol Fitó*, Jordi Guitart

Barcelona Supercomputing Center and Technical University of Catalonia, Barcelona, E-08034, Spain

ARTICLE INFO

Article history:

Received 28 October 2011

Received in revised form

6 April 2012

Accepted 17 May 2012

Available online 26 May 2012

Keywords:

Business-driven risk management

Self-managed Cloud providers

Private Cloud provisioning

Resource failures

ABSTRACT

Cloud computing is an innovative and promising paradigm that is leading to remarkable changes in the way in which hardware and software are designed and purchased, as well as how IT systems are managed. However, the Cloud is a risky paradigm. For instance, the use of Cloud services, which usually are external assets to their consumers, implies unprecedented risks that must be taken into account.

In this paper, we propose the involvement of the risk management discipline into the Cloud computing realm. We present a risk management approach led by business-level objectives (BLOs) of Cloud organizations. Its main goal is to assist in business-driven self-managed Cloud providers, by facing uncertainties always present in their internal decision-making processes. Our Cloud-aware risk management method includes a SEmi-quantitative BLO-driven Cloud Risk Assessment (SEBCRA) as the core subprocess. Its aim is to constantly rank and prioritize risks affecting the governing business-level goals.

In addition, we present, as a use case, a PaaS provider that incorporates our risk management approach to enhance the achievement of two BLOs, i.e. maximization of profit and customer satisfaction. In particular, it can manage – identify, assess, and treat – the most critical Cloud infrastructure-level risks, i.e. provisioning its private Cloud, either under- or over-provisioning, as well as resource failures. We present some risk treatment responses to face these risks and we evaluate their impact on the above-mentioned BLOs. Our results show that the best responses to address risks may change over time depending on the current provider's status. As a result, an adaptive management of risks should be considered as a mandatory process for Cloud providers to ensure their success in the ever-growing worldwide ecosystem of Clouds.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays, Cloud computing is widely recognized as the most promising computing paradigm of the last several years [1]. A recent Gartner report [2] identifies Cloud computing as the most strategic technology and trend for the majority of today's companies, basically because the use of Cloud systems leads to promising business models. The benefits for both stakeholders, i.e. providers and end-users, are actually very clear [3].

Already, several Cloud computing models have emerged: *SaaS providers*, e.g. Salesforce CRM [4], which deliver software over the Internet; *PaaS providers*, e.g. Google App Engine [5], which mainly offer virtualized execution environments to host Cloud services; and *IaaS providers*, e.g. Amazon EC2 [6], which provide virtualized computing resources as a service and, thus, serve as the foundation layer for Cloud systems. Each model offers different features and/or services, at different degrees of flexibility, and involves distinct risks. This includes new risks to be determined (due to the usage

of new technologies such as virtualization), as well as well-known risks to be re-evaluated within the Cloud domain.

The current trend involves several of those providers interoperating in Cloud ecosystems. These multi-Cloud scenarios (e.g. Cloud federations) comprise for instance PaaS providers outsourcing resources from public Clouds of third-party IaaS providers when their customers demand overcomes their private Cloud capacity. As physical Cloud infrastructures are, of course, provisioned in a static way, Cloud providers have to accurately size their private Cloud capacity in order to tighten capital expenditures (CapEx). In fact, Spellmann et al. [7] state that those initial investments are much more difficult to secure than operating expenses (OpEx), e.g. electric bills, since they are usually budgeted for technology refresh every 3–5 years. Moreover, providers are exposed to physical or virtual resource failures, which may represent significant losses for them.

All in all, Cloud providers are constantly subjected to uncertainties during their operation, which may greatly impact their business expectations. These uncertainties can represent threats for their success which, therefore, can greatly reduce the well-known benefits of using Cloud systems. In this case, the involvement of risk-aware decision-making processes into self-managed

* Corresponding author. Tel.: +34 934054282.

E-mail addresses: josep.oriol@bsc.es (J.O. Fitó), jordi.guitart@bsc.es (J. Guitart).

Cloud providers is clearly a need in order to minimize undesirable expenditures. On the other hand, uncertainties may result in opportunities or positive impacts for Cloud providers. In fact, there are events and management actions that may produce either positive or negative results for the business. For instance, overbooking resources in a Cloud provider may have a twofold impact: it increases the probability of violating service-level agreements (SLAs), and thus it implies revenue loss for the provider if violations occur, but conversely it can lead to obtaining more profit because the provider is serving more customers. Therefore, a remarkable tradeoff appears when considering the best risk-aware management action(s) to carry out in order to face risks. In fact, the best decision will highly depend on the provider's business-level objectives (BLOs) (e.g. profit and ecological efficiency maximization), which should be used to drive the whole risk management process.

1.1. Contributions

This paper, which extends our previous work [8], contributes to the inclusion of the risk management discipline into the Cloud computing paradigm.

First, we present an overview of risk management and assessment methods, as well as of the most important Cloud-related risks to be addressed.

Second, we propose the adoption by Cloud providers of a risk management process led by business interests, such as BLOs. We also propose a SEmi-quantitative BLO-driven Cloud Risk Assessment (SEBCRA) as its core subprocess. Its main goal is to constantly evaluate the impact of Cloud-specific risks, either with positive or negative consequences, on business objectives. Based on that, any Cloud organization is able to be aware, at any moment, of how to efficiently tackle uncertainties and their related risks, thus aligning its low-level management decisions with those high-level (business) objectives.

Third, we suggest several risk mitigation and adoption responses to deal with the negative and positive impacts, respectively, of the most critical infrastructure-level risks identified in the assessment step, namely the risk of provisioning a private Cloud, i.e. under- and over-provisioning, and the risk of physical and virtual resource failures.

Last, we demonstrate the applicability of the proposed BLO-driven risk management approach in a PaaS provider, which offers execution environments to host Web-based services. We experiment with different utilizations of its private Cloud (number of hosted services) and time-varying workloads for these services and we evaluate the impact of applying different risk treatment responses with respect to the achievement of two important BLOs for current Cloud providers: the maximization of profit and customer satisfaction. Our results demonstrate that a dynamic risk treatment strategy, which is able to apply the most appropriate risk response according to the provider's status to fulfill its BLOs, is needed to deal with the aforementioned Cloud infrastructure-related risks.

The remainder of the paper is organized as follows. Section 2 exposes useful background around risk management and assessment methods, as well as identifying the most important Cloud-related risks and re-evaluating traditional ones. Section 3 presents some related work on risk management and assessment. Section 4 details the business-driven risk management and assessment procedures. Section 5 presents a use case of how a PaaS provider is able to deal with critical Cloud infrastructure-related risks. Section 6 details the experimental environment and the evaluation of the presented risk management approach and of several risk treatment responses to face those risks. Finally, Section 7 draws our conclusions and exposes future work.

2. Background

2.1. Risk management and assessment

There are many different definitions of risk which have been developed and adopted by several disparate organizations over recent years. After considering dozens of them, ISO 31000:2009 [9], together with ISO/IEC Guide 73 [10], defines *risk* as the “effect of uncertainty on objectives”. It also states that risk is the consequence of an organization setting and pursuing objectives against an uncertain environment. The uncertainty comes from both internal and external events which may or not happen. In general, they can represent opportunities for benefit or threats to success, i.e. positive and negative impacts of risks on an organization's objectives. Thus, and in contrast to traditional risk avoidance (mitigation) strategies, adopting positive risks may lead to obtaining significant benefits from the business point of view. Actually, managing risks can be seen as a process of optimization that causes organizations to minimize uncertainties in achieving their objectives.

Risk management is governed by generic guidelines and principles established in the widely accepted ISO 31000:2009. By definition, it is the process whereby organizations treat, in a methodical way, risks related with their activities. Its main goal is to obtain benefits and sustainable values for the business in each of its activities and across all of them. For this reason, it should be a fundamental part of any organization's strategic management.

Risk assessment is a core subprocess of any risk management strategy. It consists of the determination of a quantitative or qualitative value of each risk, also known as risk-level estimation, related to a particular situation and a recognized event (representing either a threat, an opportunity, or both). There are three primary methods to assess risks according to [11]: *qualitative*, which roughly categorizes risks and thus does not need to determine the numerical value of all assets at risk and frequencies; *quantitative*, which assigns numerical values to both the impact and the likelihood of risks; and *semi-quantitative (or hybrid)*, which is less numerically intensive than quantitative methods and classifies (prioritizes) risks according to consequences and foreseen probabilities.

Quantitative risk assessments have been criticized for being overly reductive and diverting attention from preventive actions. In addition, they ignore qualitative differences among risks. Although the calculations involved are tedious and include a strong element of arbitrariness, their main advantage is that they provide accurate measurements of the magnitude of the impacts. However, those quantitative impacts may be unclear, thus requiring to be somehow interpreted in a qualitative way. In contrast, the main advantages of qualitative assessments are the prioritization of risks and the identification of the most important areas for improvement. Even so, they do not provide enough quantifiable measurements concerning the probabilities and impacts of risks. As a result, semi-quantitative methods basically take advantage of both these aforesaid aspects and, therefore provide risk prioritizations and useful (semi-)quantifiable impact analyses.

2.2. Is risk management needed in Cloud organizations?

New risks have appeared together with the evolution of the Cloud computing paradigm. Within them we find specific issues imposed by law or regulations, as well as operational risks inherent to the use of Cloud systems, either local or external assets. These risks can have a great impact on the operation of Cloud providers, making it inconsistent with their respective business strategies, represented by means of business objectives (BLOs) and/or constraints. As suggested by ISO 31000:2009, proper

risk management strategies driven by business aspects must be integrated into a Cloud organization's decision-making processes in order to be effective. Here is where and how uncertainties (risks) are generated and thus can be tackled. All the Cloud entities, not just the providers as a whole, should be the subject of risk management and assessment approaches. In fact, risks in Cloud systems must be considered at service, data, and infrastructure layers. Notice that the level of risk will, in many cases, vary significantly with the type of Cloud environment being considered. For instance, being a participant of a federation of Clouds involves more risks than only managing a private Cloud. Risk management strategies should be able to clearly identify new Cloud-specific risks and re-evaluate well-known risks in a Cloud environment.

Cloud-related risks. There is an ongoing effort to identify the risks associated with the Cloud computing paradigm [12–16]. As a result, we can note the following risks that Cloud computing introduces: risks associated with the virtualization technology, such as failures in multi-tenancy, virtual machine (VM) isolation, and hypervisor vulnerabilities; loss of direct control of resources and software such as provider lock-in, decreased reliability since providers may go out of business, agreement (SLA) breaches; risks associated with data such as data protection responsibility, insecure or incomplete data deletion; and legal risks such as regulatory compliance, data location, and the effect of international boundaries on operation.

Re-evaluating traditional risks in Cloud environments. The impact of some traditional risks must be also re-evaluated in Cloud environments. For instance, the risk of network partitions is now more critical for Cloud organizations since they are totally based on the network. Other risks, such as natural disasters, should be considered in a different way – because of the constant use of external resources – to ensure high availability of Cloud services.

In short, Cloud organizations need to definitely include risk management processes, as well as develop risk-aware strategies, policies, and heuristics required to face new Cloud-specific risks and traditional ones that have changed within this new scenario.

3. Related work

The risk itself and its management have been considered in a great number of research fields, e.g. statistics, biology, engineering, and systems analysis, for many years. In this regard, Aven has made pioneering contributions for risk analysis and management, such as [17]. He has introduced the concept of considering a risk as an event where the result is uncertain, either positive and negative [18]; and he has stated that semi-quantitative analysis replaces tedious quantitative approaches very well [19]. Generally, the research contributions toward risk management and assessment differ in the level of development of methodology items. According to this, they can be grouped under the three risk assessment methods: quantitative [20], qualitative [21,22], and semi-quantitative [19,23,24].

Apart from these general risk-related publications, the management of risks has been widely tackled in other successful distributed computing paradigm, i.e. Grid computing. In this field, and within the context of the AssessGrid project [25], Djemame et al. have contributed with a lot of research works [26–29]. They have widely addressed the inclusion and implementation of risk management and assessment methods into Grid environments. They treat risk-aware negotiations and SLAs, risk-based decision-support for infrastructure management, and calculation of risk-indicators for providers ranking and competition, among others. Whereas we share a common interest of introducing risk management in distributed systems, we need to focus on risk-aware decision-making policies aimed at managing virtualized (Cloud)

infrastructures and services running on top of it that can take into account the specifics of these kinds of environment. Moreover, Voss et al. [30–32] present a Grid risk management process which is based on the standard for risk management processes developed by the Federation of European Risk Management Associations (FERMA) [33]. They address the risk analysis during SLA negotiations, implement a risk-aware scheduling, and consider job migrations in a risk-aware concept. Note that the risk management approach proposed herein is also based on the FERMA standard.

Yeo and Buyya [34] pose the problem whether a resource management policy implemented in the commercial computing service is capable of meeting the required objectives or not. They develop two evaluation methods to validate the effectiveness of resource management policies in achieving the required objectives: separate and integrated risk analysis. The experimentation performed demonstrates the applicability and success of their risk-based methods. However, they only consider one business objective, i.e. profit maximization. This objective is also taken into account individually in [35,36], where the risk of paying penalties to compensate service providers' users is minimized. Thus, they are able to increase the profit of providers. In addition, the effects of the risks of resource failures and SLA compliance are analyzed in [37]. The authors evaluate through simulation how users and providers can minimize their costs by adding penalties to cover the risk of not fulfilling SLAs. We are going further these works by considering multiple business objectives at the same time. This opens several research issues yet to be solved.

4. BLO-driven Cloud risk management

In this section, we describe a Cloud risk management process driven by an organization's high-level interests (i.e. BLOs). It is designed to address the impacts and consequences of Cloud-related risks into well-defined BLOs of a given Cloud organization, such as maximizing customer satisfaction or profit.

Our risk management approach for Cloud organizations, which is depicted in Fig. 1, is based on the risk management process described in ISO 31000:2009. It is governed by the organization's BLOs and strategic objectives, although it can also consider IT-level performance objectives that contribute to BLO fulfillment, such as maximizing reliability or minimizing energy consumption.

The risk management process is executed when the management policy of the Cloud provider needs to actuate the system. This can occur when either a user asks for an action to be performed (e.g. a new service must be deployed), an IT-level event takes place (e.g. a resource failure is detected), or periodically. The management policy will use the risk management process to assess the risks and decide the best action to fulfill the provider's BLOs. Note that at each step new risks can arise, some of them can disappear, or the impacts could be different.

We can distinguish the following subprocesses: *SEBCRA* (*Risk Assessment*), which is the overall process of risk identification, analysis and evaluation (see Section 4.1 for a detailed explanation); *Risk Treatment*, which implements and selects risk-aware responses, as well as measures, actions, and controls to face risks; *Risk Monitoring*, where all the above steps are reviewed; and *Risk Reporting* and communication.

4.1. SEmi-quantitative BLO-oriented Cloud Risk Assessment

Risk management literature commonly specifies the need to rank and prioritize risks in order to identify areas for immediate improvement and, thus, focus the best efforts on minimizing and maximizing the effect of threatening events and opportunities, respectively. With this goal in mind, we present a new risk assessment model, i.e. a SEmi-quantitative BLO-oriented Cloud

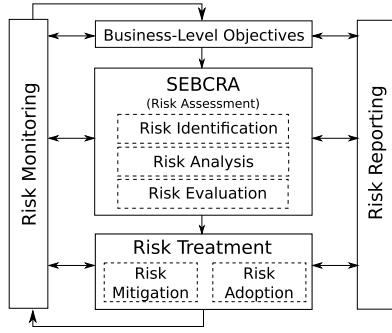


Fig. 1. BLO-driven Cloud risk management steps.

Risk Assessment (SEBCRA), which has the purpose of constantly ranking Cloud risks. Its main difference compared other risk assessment models is that it evaluates the impact of risks on an organization's global BLOs, instead of considering the effect on different single assets of the organization.

The risk assessment is the core process of the BLO-driven Cloud risk management, and it provides risk-level estimations (RLEs) as outputs, which are individually specified for each risk and BLO affected. Generally, the risk assessment is subdivided into identification, analysis, and evaluation of risks.

4.1.1. Risk identification

This first step establishes and defines an organization's potential events, uncertainties and, thus, risks. It is perhaps the most difficult aspect of managing risks, because more events will happen in the future than can be predicted today, i.e. the problem of "unk-unks" (unknown unknowns) [38]. In addition to the identification of risky events, this stage describes them in a structured format, e.g. by using a table, in order to guarantee a comprehensive risk assessment method. Within this integrated description we can highlight some fields, such as the scope of the event in question, its quantification, assets, and resources at risk, as well as tolerance and potential actions for improvement.

4.1.2. Risk analysis

Based on each risk previously identified, this step is aimed at figuring out the events' likelihood of occurrence and the magnitude of their consequences, i.e. the estimated impact on BLOs. We propose a semi-quantitative risk analysis, which uses a risk-level matrix, as suggested in ISO/IEC 27005:2008 [39], in order to carry out risk-level estimations. This ISO has become a widely accepted guide for information security risk management. It provides meaningful insights on how risk analysis and treatment processes should be developed within the context of all types of organization. In particular, the impact is usually evaluated using a 3×3 , 4×4 or 5×5 risk-level matrix [13], depending on the granularity desired, which only considers the negative side of the risks. Nevertheless, we use a 10×5 matrix because we are considering the impact in terms of threats (downside risks) and opportunities (upside risks).

Going into detail, we establish the following semi-quantitative classifications: the *probability* of occurrence of a risky event (P_i), expressed by means of very unlikely (1, e.g. once in 20 years), unlikely (2, e.g. yearly), possible (3, e.g. monthly or weekly), likely (4, e.g. daily), and frequent (5, e.g. at any moment); the *impact* of such event (I_i), either a threat, a benefit, or both, semi-quantified between very high (-5 or 5, for negative and positive impact, respectively), high (-4/4), medium (-3/3), low (-2/2) and very low (-1/1); the *BLO affected* (B_i); and the *risk-level estimation* ($RLE_i(B_i)$), which is proportional to the *probability* of a given event (P_i) and

its *impact* on the BLO in question ($I_i(B_i)$), resulting in the following equation:

$$RLE_i(B_i) = P_i \cdot I_i(B_i).$$

In this case, five levels of RLE are defined: *critical* if $-25 \leq RLE \leq -15$; *unacceptable* if $-15 < RLE < -5$; *negligible* if $-5 \leq RLE \leq 5$; *profitable* if $5 < RLE < 15$; and *highly profitable* if $15 \leq RLE \leq 25$. The risk analysis results (i.e. RLEs) can be used to produce risk profiles. These are very useful for ranking risky events and, thus, for prioritizing risk treatment efforts.

For a better understanding, Table 1 illustrates all the possibilities concerning RLEs for a given BLO in terms of semi-quantitative ranges. Note that values given above to qualitative measures of probabilities and impacts are only aimed at defining those different ranges of risk level estimations, as well as assessing them.

4.1.3. Risk evaluation

This step compares the estimated risk levels against a risk acceptance criterion, which is a threshold established by business executives. In fact, the action of moving up and down this threshold is a serviceable practice to determine the organization's acceptable risk level. This step results in the determination of whether each risk should be treated or not.

4.2. Risk treatment: retainment, mitigation, or adoption

Once risky events has been identified, analyzed, and evaluated, the risk treatment subprocess takes place. In a nutshell, the main goal is to treat events in order to minimize threats and maximize opportunities. This involves defining new or improving existing risk-aware responses, controls, and management actions to conduct appropriate risk treatment strategies.

In general, this is a crucial step when managing risks, since the future impact of any event may vastly depend on the risk treatment strategies selected. Note that the best responses to address risks may change over time depending on the current provider's status, e.g. the utilization of its in-house infrastructure or workload received. This stage includes predicting the resulting risk-level estimations for each potential risk response that can be applied. Using these estimations, the provider can choose the risk response that is better aligned with its BLOs, given its current status.

As stated before, day-to-day uncertainties present in any organization can represent both positive and negative impacts for businesses. Traditionally, strategies for dealing with risks representing threats for the business have been proposed. However, it is worth considering also how to exploit events that may represent opportunities. For instance, the number of clients could be higher than expected, which could lead to higher revenue if we ensure that we can attend to all of them. Similarly, some client could finish its execution earlier than expected, which could also lead to higher revenue if we use the freed resources to attend to other clients. We distinguish then among two types of strategy, depending on the nature of the risks: mitigation and adoption strategies for dealing with negative (threats) and positive (opportunities) events, respectively.

Negative risks (threats). Risk mitigation strategies include, among others, the reduction of the probability that a threatening event will occur and/or the reduction of its negative effect if it occurs. Generally, it is aimed at improving negative risk-level estimations. In particular, there are three possible risk mitigation responses to effectively deal with threats.

- *Avoid* the risk, by eliminating its root cause(s) or changing the organization's business plan.

Table 1
Risk-level matrix (in qualitative ranges) of SEBCRA.

		Probability P_i				
		Very unlikely (1)	Unlikely (2)	Possible (3)	Likely (4)	Frequent (5)
Impact $I_i (B_i)$	Benefit	Very high (5)	Negligible	Profitable	Highly profitable	Highly profitable
		High (4)	Negligible	Profitable	Highly profitable	Highly profitable
		Medium (3)	Negligible	Profitable	Highly profitable	Highly profitable
		Low (2)	Negligible	Profitable	Profitable	Profitable
		Very low (1)	Negligible	Negligible	Negligible	Negligible
	Threat	Very low (-1)	Negligible	Negligible	Negligible	Negligible
		Low (-2)	Negligible	Unacceptable	Unacceptable	Unacceptable
		Medium (-3)	Negligible	Unacceptable	Unacceptable	Critical
		High (-4)	Negligible	Unacceptable	Critical	Critical
		Very high (-5)	Negligible	Critical	Critical	Critical

- *Reduce* the risk by taking steps to cut down its probability, its impact, or both at the same time. This is the most common approach in risk mitigation strategies.
- *Transfer* or delegate part or all of the negative effects of a threat to third-party organizations. Usually, this response incurs a cost associated with insuring the negative risk.

Positive risks (opportunities). Risk adoption strategies are aimed at taking advantage of events which may represent a benefit or positive impact for a provider's business objectives. In this case the aim is to enhance the result of positive risk level estimations. Here one can also differentiate between three possible responses for adopting risks.

- *Exploit* the risk, trying to ensure that everything is in place to make the positive risk happen.
- *Enhance* the risk, by increasing its probability and/or impact.
- *Share* the risk, which means sharing responsibility with another organization in order to exploit a risk. This is a very useful response when this exploitation is not possible without collaboration.

Negligible risks. In addition, there is a common risk response to both types of risk: *Retain* the risk, which involves accepting its consequences. This is a very common and suitable strategy for small or negligible risks, where the cost of applying any other risk response, aimed either at mitigating or adopting the risk, is greater than the potential gains. In addition, other times there is nothing to do about the risk, regardless of the probability and impact of the event causing it.

4.3. Risk monitoring

This step reviews and evaluates the effects of any risk treatment strategy being applied. In general, it assists the organization in knowing if risk-aware responses and actions are suitable or not when managing each risk. This is very important for self-management in Cloud providers driven by BLOs, since they are constantly subjected to unexpected changes in their Cloud environments. In this sense, this step ensures that appropriate risk treatment strategies are being carried out in cases of (1) existing risks changing and (2) new ones emerging as a consequence of applying a particular risk treatment response or strategy. For instance, if a provider decides to accept the risk of deploying a new service on its private Cloud infrastructure, then several new risky events emerge during the operation of such service, such as the possibility of violating its SLA. This step produces outcomes that are useful for adapting any of the previous subprocesses within the management of risks, such as the outcomes of the SEBCRA procedure. In addition, it is useful for business executives in order to rethink high-level objectives according to the constantly improved knowledge about risks.

4.4. Risk reporting

The risk management approach has to report externally all the results obtained from managing risks. Actually, these reports are provided to fulfill mandatory requirements related to risk management. However, this step is out of the scope of this paper.

5. Use case: business-driven management of infrastructure-level risks in a PaaS provider

In this section, we exemplify the suitability of the BLO-driven risk management procedure for Cloud organizations described above. We present, as a use case, a PaaS provider (from now on denoted as PaaS) which offers execution environments to host Web-based services. The aim is to demonstrate how the management of risks is able to assist decision-making processes within Cloud providers in improving the achievement of their BLOs. In this sense, we propose several risk treatment responses to face the most critical infrastructure-level risks, namely the risks of over- and under-provisioning its private Cloud, and the risks of resource failure, whether physical or virtual. The evaluation of those risk responses can be found in Section 6.2.

Multi-Cloud scenario. We consider a scenario in which PaaS providers interoperate with external IaaS providers. In this context, a PaaS may be responsible for operating, maintaining, and managing its private Cloud, which is composed by its in-house virtualized resources, and is able to outsource resources to public Clouds managed by underlying third-party IaaS providers. In general, outsourcing deals take place when the private Cloud capacity is insufficient to support the workload.

Risks and BLOs of the PaaS. First, we identify the most important risks the provider is exposed to in the scenario considered. In particular, risks are classified into the following groups, depending on the source of the causing event.

1. *Infrastructure*, including risks such as wrong capacity provisioning (both over- and under-provisioning a private Cloud), failures of physical and virtual resources, virtual machine isolation, and virtualization performance overhead, to name a few. In addition, here we group the risks of power loss to the IT systems, natural disasters, fire, etc.
2. *Services*, including risks such as accepting new SLAs, their violations due to poor performance or service disruptions, loss of governance, and hidden costs for outsourced services, etc.
3. *Data*, including the risks of data integrity loss, destruction of data, etc.

Second, we identify the business objectives that drive our risk management strategy. Examples of potential BLOs for Cloud providers are the maximization of profit (*ProfMax*), customer satisfaction (*SatMax*), and reputation (*RepMax*). We also consider IT-level performance objectives that contribute to the fulfillment of the BLOs, such as the maximization of quality of service (*QoSMax*), reliability (*RelMax*), and energy efficiency (*EnEffMax*). Obviously, other high-level objectives and constraints can be defined.

Table 2

Example of assessing the uncertain effects on potential PaaS's BLOs of inadequately provisioning a private Cloud.

Event i	P_i	B_i	$I_i(B_i)$		$RLE_i(B_i)$	
			Benefit	Threat	Benefit	Threat
Over-provisioned	Frequent	ProfMax	-	High	-	Critical
		RelMax	Medium	-	Highly profitable	-
		RepMax	Medium	-	Highly profitable	-
		QoSMax	Medium	-	Highly profitable	-
		SatMax	Medium	-	Highly profitable	-
		EnEffMax	-	Very high	-	Critical
Under-provisioned	Frequent	ProfMax	Medium	Very high	Highly profitable	Critical
		RelMax	-	High	-	Critical
		RepMax	-	High	-	Critical
		QoSMax	-	High	-	Critical
		SatMax	-	High	-	Critical
		EnEffMax	Low	Low	Profitable	Unacceptable

Table 3

Example of using the SEBCRA procedure for assessing uncertain effects of several Cloud-related events on the profit maximization BLO.

Event $_i$	Probability (P_i)	$I_i(B_i)$		$RLE_i(B_i)$
		Benefit	Threat	
Over-provisioned infrastructure	Frequent	0	High	Critical
Under-provisioned infrastructure	Frequent	Medium	Very high	Highly profitable / Critical
Physical or virtual resource failures	Possible	0	Very high	Critical
Request for new Cloud service	Likely	High	Low	Highly profitable / Unacceptable
SLA violations	Likely	Very low	Medium	Negligible / Unacceptable
Service disruptions	Possible	0	Medium	Unacceptable
Loss of service performance	Possible	0	Medium	Unacceptable
Outsourcing hidden costs	Unlikely	0	High	Unacceptable
Failure in VM isolation	Very unlikely	0	High	Negligible
Virtualization performance overhead	Very unlikely	0	High	Negligible
Data integrity loss	Very unlikely	0	Medium	Negligible
Destruction of data	Very unlikely	0	Medium	Negligible
Loss of governance of outsourced services	Very unlikely	0	Medium	Negligible
Power loss of IT systems	Very unlikely	0	Very low	Negligible
Natural disasters, fire, etc.	Very unlikely	0	Very low	Negligible

5.1. SEBCRA for risk assessment in a PaaS provider

In this section, we demonstrate the feasibility of the SEBCRA procedure to assist risk-driven self-management in Cloud providers owning in-house virtualized infrastructures. We present, as an example, how different risks have distinct impacts on BLOs. In particular, the risk of inadequately provisioning a private Cloud (either under- or over-provisioning) may have the impacts and risk level estimations on BLOs as given in Table 2. Note that this assessment is only an example, and that the values given can change depending on both the provider's status and risk treatment strategies being implemented.

On the one hand, the risk of over-provisioning can appear at any moment (probability of occurrence equal to frequent) and its risk levels estimations are *critical* for *EnEffMax* and *ProfMax*, because the provider is consuming more energy than the strictly needed, and it is paying for more resources than necessary, respectively, and *highly profitable* for *RelMax*, *RepMax*, *QoSMax* and *SatMax*, since the provider's reliability and reputation, QoS offered and customer satisfaction can be partially or totally assured if there are more resources than those needed by the services hosted.

On the other hand, the risk of under-provisioning has the same probability (frequent), but dissimilar impacts, ranging from *highly profitable* to *critical* for *ProfMax*, because the provider can outsource resources when needed and thus improve its overall profit earned, but penalties for SLA violations can be significant; *critical* for *QoSMax*, *RelMax*, *RepMax* and *SatMax*, because the provider may be not able to meet the QoS agreed in the SLA, and this inappropriate QoS delivery leads to a decrease of clients' satisfaction, as well as the provider's reliability and reputation; and ranging from *profitable* to *unacceptable* for *EnEffMax*, because, although its energy consumption can be significantly lower, the

provider should outsource resources in the case of applying the under-provisioning technique. For this reason, it might not incur significant improvements in terms of energy efficiency since the outsourced resources may consume even more energy than if they were to run in local nodes.

5.2. SEBCRA for profit maximization in a PaaS provider

In this section, we demonstrate how the SEBCRA procedure can be very useful for determining ways of improving a given BLO, which is profit maximization. Note that risks can have an impact on many BLOs (as shown in Table 2), but in this section we only present an example of the consequences on the *ProfMax* BLO. The key parameters of the SEBCRA procedure for each risk affecting the *ProfMax* BLO are shown in Table 3. The different risks have been derived from the analysis of the work in the area, which we introduced in Section 2.2. Note that some events could be a consequence of others, but they are different risks with distinct impacts. For instance, the risk of service disruption is usually due to resource failures, but their effects on BLOs could be different.

After completing the table of probabilities and impacts, the SEBCRA procedure helps the PaaSP in the tasks of categorizing and prioritizing risks according to their importance (impact) for a given BLO. In this sense, the provider is able to constantly put its best efforts into addressing threatening events (or exploiting opportunities) that may incur more losses (or benefit) to the *ProfMax* BLO. In this example, this occurs with the risks related with the Cloud infrastructure, namely the risks of over- and under-provisioning its private Cloud and of physical and virtual resource failures.

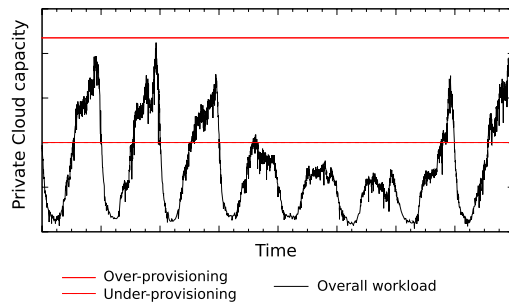


Fig. 2. Private Cloud provisioning strategies.

5.3. Treating the risks of provisioning a private Cloud

Once the risk assessment has been carried out, one needs to treat risks exceeding the acceptance threshold determined by the business entity. To that end, a suitable risk treatment process needs to propose risk-aware responses and actions which must help organizations to mitigate or exploit risks with negative and positive consequences, respectively. Note that not all possible risk responses (those briefly described above in Section 4.2) can be applied to face each risk. In fact, the risk of over-provisioning, for instance, cannot be directly avoided. The “avoid” risk response would lead to the provision of an infrastructure with fewer resources, which could imply switching to the risk of under-provisioning.

As shown in Fig. 2, an over-provisioned infrastructure implies that most servers are underutilized in low demand situations, with the corresponding energetic, economic, and administration expenditures. On the other hand, an under-provisioned private Cloud means that the provider will not pay so much for these costs. However, it will lose some of the clients as it is not able to cope with peak demands. In addition, although a Cloud provider assumes that it has the required capacity for meeting peak demands, a wrong estimation can be made or some unexpected demands due to sudden events could appear. Several risk treatment responses to deal with these risks of wrongly provisioning a Cloud infrastructure have been identified.

Note that, in addition to these responses, we also include the option (*Retain*) of just accepting these risks. In this case, we assume the use of a round-robin scheduler. Therefore, services are deployed in a way that nodes are progressively filled, thus trying to ensure an equitable distribution of VMs among physical resources of the private Cloud.

Risk of over-provisioning. We propose the following risk mitigation responses to deal with the risk of over-provisioning a private Cloud.

- **Reduce 1:** this both consolidates services’ VMs and applies turning on/off of nodes depending on the workload handled by the provider’s infrastructure. In other words, it aligns the number of running nodes with the workload received, while unused ones are powered off. Thanks to virtualization, several service components (VMs) can be running on the same physical host in isolation with low performance impact. This consolidation of resource pools is also a very common practice for minimizing the energy consumption of data centers [40], and thus costs. Recent studies, such as [41], have shown that on average an idle server consumes roughly 70% of the power consumed by the same server running at the full CPU speed.
- **Reduce 2:** apart from consolidating VMs and turning on/off nodes, this response is based on overbooking physical nodes to reduce even more the number of running nodes. An overbooking factor guides the placement policy. In particular, we use 1.5 and 1.25 as high and low overbooking factors, respectively.

One could think also about *Sharing* this risk with external providers by insourcing resources to them. This means offering a set of internal resources to let them, for instance, extend their private Cloud with external resources. However, the applicability of this response depends on whether the external demand for these resources is higher than the cost of keeping these resources online or not. For this reason, we do not include this response in our use case.

Risk of under-provisioning. In this case, two risk treatment responses are suggested (apart from retaining the risk).

- **Avoid:** this rejects the deployment of new services if there is not enough capacity in the private Cloud to support their operation.
- **Share:** this tries to ensure the potential positive impact of this risk already identified in Table 3. In this sense, the provider decides to share the responsibility of this risk with external IaaS providers, by outsourcing the deployment and operation of services. In general, outsourcing deals are very suitable when the PaaS needs to scale up its limited in-house infrastructure in order to meet peak demands.

Note that the first risk treatment response aims at mitigating the negative consequences of this risk (*Avoid*), whereas the second tries to ensure that the opportunity is realized (*Share*). In addition, the latter represents a possibility of avoiding the threatening part of this risk, since the provider should be able to meet resource requirements of services.

5.4. Mitigating the risk of resource failures

Unexpected failures of physical and virtual resources are becoming a usual undesired phenomenon in current data centers. This fact is leading Cloud providers to take a deep look at best practices to tackle the threatening risk of such failures, and their associated uncertain effect on the provider’s business objectives. In this sense, the exploration of suitable risk treatment strategies seems one of the possible ways to face them. We propose the following four possible risk mitigation responses.

- **Retain** the risk and its consequences. This can be very dangerous for Cloud providers, since service outages represent significant losses in terms of profit, reputation, reliability, and customer satisfaction. However, it can be appropriate if the mean time between failures (MTBF) is very large.
- **Avoid** the risk, by replicating virtual resources supporting the operation of services. This response guarantees that there is not a single point of failure, but the provider’s operational costs are increased.
- **Reduce** the consequences, by recovering service execution upon resource failures. With this response, the provider must ensure the minimum mean time to repair (MTTR) to meet the service availability agreed with customers.
- **Transfer**, by outsourcing the operation of preferential services to third-party IaaS providers. Agreed SLAs with those providers should ensure a reward in case of unavailability of outsourced services.

Note that both *Avoid* and *Reduce* responses also outsource the service operations to external IaaS providers if the private Cloud does not have enough capacity to cope with the load received by the provider. In addition, all responses consolidate VMs and turn on/off physical nodes depending on the demand.

6. Experimentation

In this section, we evaluate the impact in the PaaS’s BLOs when incorporating the proposed risk treatment strategies into

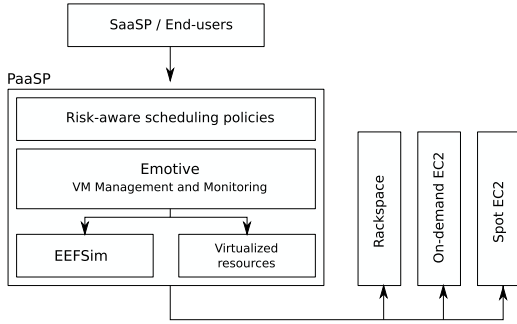


Fig. 3. Experimental environment.

its self-managed operation. In particular, we have tested how such a provider is able to fulfill two BLOs – maximization of profit and customer satisfaction – when facing what the SEBCRA procedure has identified as the most critical Cloud infrastructure-related risks, i.e. risks of provisioning a private Cloud and risks of resource failures. With this goal in mind, we have experimented with the PaaS in distinct situations, i.e. with different workloads and customer demands (number of services hosted), and we have measured the fulfillment of these two BLOs achieved by the different risk treatment responses detailed above.

6.1. Experimental environment

The experimental scenario, which is depicted in Fig. 3, is composed of different stakeholders: (1) end-users, who want to use services deployed on the Cloud; (2) SaaS providers, who want to deploy their services on the Cloud; (3) the PaaS provider (PaaS), which is a Cloud hosting provider with its own private Cloud that offers execution environments to host Cloud services (by means of the EMOTIVE Cloud middleware [42]); and (4) three third-party IaaS providers to which the PaaS can outsource the operation of services to their public Clouds. Note that all risk treatment responses evaluated in this section have been developed as scheduling policies for the EMOTIVE Cloud middleware.

6.1.1. EMOTIVE and EEFSim simulator

The EMOTIVE middleware allows executing Cloud services and provides customized virtual environments to them. In addition, it allows managing Cloud providers according to different policies. EMOTIVE can be used on either a real Cloud environment or a simulated one. This experimentation is centered on the latter case, in which the EEFSim simulator [43] reproduces the behavior of a real virtualized data center. The use of simulators has offered remarkable advantages in research, and even more in the Cloud computing area where we work with large-scale distributed environments.

6.1.2. Experimentation parameters

In the experiments, the PaaS has a private Cloud composed of 50 physical nodes. All of them use Xen as the VM hypervisor.

Energy. The PaaS is located in UK, so we use the corresponding power pricing [44], i.e. 0.101€/KWh, for using the common grid power. The EEFSim simulator provides the energy consumption of the private Cloud. It follows a Gaussian distribution $\mathcal{N}(PC, 1)$, where PC is the real power consumption previously obtained with a power meter (see [45] for further information on the power usage characterization used). Note that a Power Usage Effectiveness (PUE) of 2.0 – this is the industry average value according to [46] – is used for calculating the real energy consumption of the PaaS's datacenter.

Operational incomes and costs. The PaaS offers three different prices to clients (i.e. SaaS providers) for the hosting of services, depending on the level of QoS desired: 0.5€/h, 0.35€/h, and 0.2€/h for *Gold*, *Silver*, and *Bronze* quality, respectively. This categorization is very useful for scheduling policies responsible of allocating resources to services, or selecting the best ones to be canceled or outsourced. In addition, the PaaS has three options for outsourcing resources: *Rackspace Hosting* [47], at a cost of 0.4€/h; and *Amazon EC2* [6], at a cost of 0.337€/h or 0.162€/h for on-demand and spot large standard instances, respectively. These prices comprise the costs of computation (2 CPUs), storage (2 GB), and bandwidth (5 GB of data transfer in and out). With this amount resources the PaaS ensures that the SLAs of outsourced services will never be violated, regardless of resource utilization over time.

The PaaS's operational expenditures are equal to the price of the local and external infrastructures used. The former includes the amortization costs of local hosts (over 4 years), the space (over 10 years) required to deploy them using a price of 2000€/m², and the electricity consumed in terms of the annual billing. The latter comprises only the costs of outsourcing VMs in terms of usage per hour, i.e. following the pay-as-you-go billing model.

SLA violation penalties. The PaaS is subjected to suffer losses if SLAs agreed with its clients are not met. We experimented with a minimum availability of four nines (99.99%). Service availabilities below this threshold are considered as SLA violations. In addition, an SLA i is considered as violated if the response time offered to clients is greater than a threshold $MVRT_i$ (minimum violating response time). We used a value of 8 s, since this is the limit time a user is supposed to wait for a Web page according to [48]. As shown in Fig. 4, penalties associated to such SLA unfulfillment are assessed by two parameters: (1) the *Misdelivery*, namely the deviation from the $MVRT_i$, which follows a linear function; and (2) the *Unavailability*, namely the total amount of time that the service is unavailable, which is determined by a Gompertz function (a kind of sigmoid function). The latter is a type of mathematical model for time series, where growth is slowest at the beginning and end of a time period. SLA penalties can be modeled in a realistic way by using this function, because it allows to reduce penalties for low SLA violations, while setting also an upper limit for high ones. In order to be coherent, each QoS level has a different increase in these functions. The amount of penalty is given by $\%Penalty = Price \cdot \frac{(Unavailability + Misdelivery)}{100}$. We have set a maximum possible value for this penalty equal to 75%, 100%, or 125% of the price paid by clients, depending on the QoS agreed. We guarantee enough compensation to customers if the service offered is far below of what was agreed in the SLA. However, we apply this upper bound in order to avoid malicious users, who can overload services themselves with the aim of receiving large amounts of monetary rewards.

6.1.3. Workload and Cloud services

The Cloud services hosted in the PaaS receive requests following a workload pattern obtained from an anonymous European ISP (collected during 2009). The workload pattern used is of a whole week, thus representing the typical time-varying users demand over this period. Actually, we extracted each individual service's workloads, which are used in the experimentation to reproduce real accesses to different services over the simulated time. In addition, note that those services have different computing resource requirements. We carried out several tests, each one simulating one week and with different amount of services running in the provider. In any case, the aim is to demonstrate the success of the business-driven risk management approach presented, regardless of the number of services hosted and, therefore, of the private Cloud's utilization.

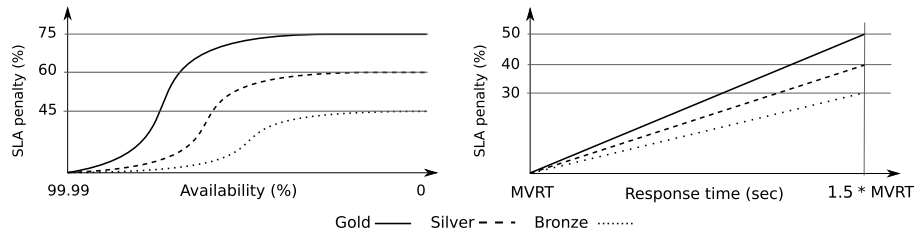


Fig. 4. Different assessments of SLA violation penalties.

The service application to be hosted in the PaaSP is the SPECweb2009 e-Commerce [49]. Its performance model has been incorporated into the simulator. It was obtained by performing several real tests, on a Tomcat server, facing different input loads and with different processing units (from 0.25 to 4) available to it (note that such application is CPU bounded in this environment). Indeed, such modeling is focused on the response time metric. A performance pattern which relates this metric with both incoming users' load and CPU usage of the server has been detected.

6.1.4. Simulating a faulty Cloud environment

The reliability of Cloud providers is threatened by unexpected resource failures which imply negative impacts in several ways. Among others, the availability of services hosted in faulty infrastructures is clearly affected. We consider that a resource, either physical or virtual, has failed if there is a crash failure, i.e. the resource halts, or an omission failure, i.e. the resource does not answer to petitions. By definition, the failure rate (λ) is the frequency with which a given system fails, expressed in failures per hour. In practice, the mean time between failures (MTBF, $\frac{1}{\lambda}$) and mean time to repair (MTTR) are often used. The former is the average time between failures, while the latter determines the time taken to repair a failed resource, either physical or virtual. These parameters are commonly used to assess the availability as follows:

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}.$$

Availability is one of the most common service level objectives (SLOs) present in SLAs offered by current commercial providers. For instance, Amazon EC2 ensures in its SLA [50] an annual uptime percentage of at least 99.95%. Taking into account a minimum MTTR of 15 min (manual rebooting), the maximum MTBF should be roughly 21 days. However, this is not real according to real outages till 2008. Apart from that, Google reported an average of 1000 node failures/year in their typical cluster of 1800 nodes, which represents an MTBF of 8.67 h [51]. This is actually the value used in this experimentation. With regard to failures of virtual machines, we adopt the MTBF used in [52], i.e. 7.5 days. Finally, and as stated above, we use an MTTR of 15 min.

6.2. Maximizing the PaaSP's profit and customer satisfaction

This section demonstrates the applicability of the business-driven risk management approach proposed to deal with threatening infrastructure-level risks to which Cloud providers are exposed. We analyze how different risk mitigation responses are capable of shifting the consequences of those risks from negative to positive, depending on the current provider's status. We evaluate the impact of those risks on two BLOs: maximization of profit (*ProfMax*) and customer satisfaction (*SatMax*). The profit is expressed in €/week and it is calculated using the typical cost-benefit analysis: $\text{Profit} = \text{Revenue} - (\text{Operational costs} + \text{Penalties})$, while customer satisfaction is assessed as the percentage of fulfillment of SLAs agreed with them.

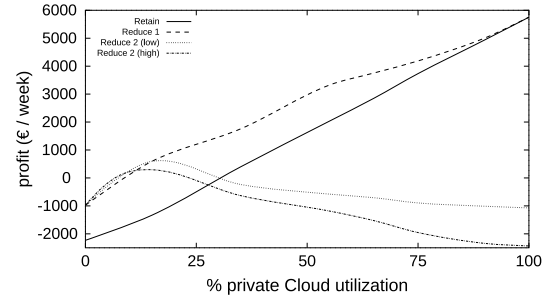


Fig. 5. The effects of over-provisioning the private Cloud on a provider's weekly profit, depending on both risk treatment response used (described in Section 5.3) and the number of services hosted.

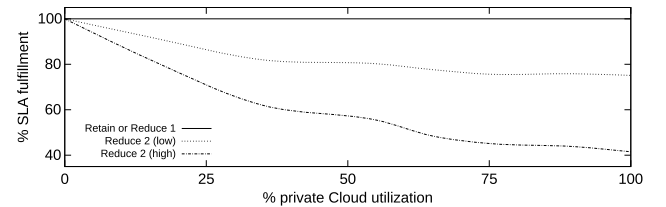


Fig. 6. The effects of over-provisioning the private Cloud on customer satisfaction, expressed as a percentage of SLA fulfillment, depending on both the risk treatment response used (described in Section 5.3) and the number of services hosted.

6.2.1. Risks of provisioning the private Cloud

We distinguish between two types of risk at this level, which are over- and under-provisioning physical infrastructures. In other words, these are the risks of building a private Cloud with a wrong capacity, either insufficient or more than necessary. In our experimental setup, considering both the amount of resources (capacity) available in the private Cloud and the services' demand in terms of computing resources, we can calculate the number of services that can be executed with that capacity. Given that the PaaSP has 50 physical nodes, it can execute up to 140 services. According to this, the PaaSP is under- or over-provisioned if there are more than or fewer than 140 services, respectively.

Over-provisioning. Figs. 5 and 6 illustrate the effects of over-provisioning a private Cloud on the PaaSP's weekly profit and its customer satisfaction, respectively. The results are presented for each different risk mitigation strategy (those described in Section 5.3), and depending on the utilization of the private Cloud (i.e. the number of services hosted).

As shown in Fig. 5, the worst risk response in terms of *ProfMax* BLO fulfillment is *Reduce 2*, either with a 1.5 (*high*) or 1.25 (*low*) overbooking factor. This response overbooks physical resources to reduce the amount of online hosts and, thus, reduce the energy bill. However, the number of penalties due to SLA violations is significantly higher. In fact, the fulfillment of the *SatMax* BLO is also clearly diminished (Fig. 6). Altogether, *Reduce 2* response, which overbooks resources according to a factor, is only appropriate if the private Cloud utilization is lower than 15%. Response *Reduce 1* is the best in terms of achieving the maximization of both BLOs. It consolidates several services' VMs on each physical host, but

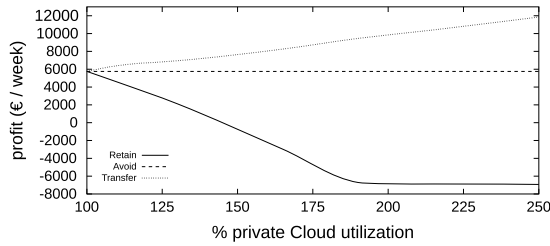


Fig. 7. The effects of under-provisioning the private Cloud on provider's weekly profit, depending on both the risk treatment response used (described in Section 5.3) and the number of services hosted.

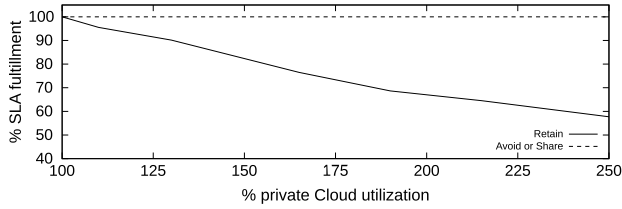


Fig. 8. The effects of under-provisioning the private Cloud on customer satisfaction, expressed as a percentage of SLA fulfillment, depending on both the risk treatment response used (described in Section 5.3) and the number of services hosted.

ensuring that the services have the necessary computing resources. This implies both significant economic savings in the energy billing and total fulfillment of SLAs agreed with customers. The *Retain* response, which uses a round-robin scheduler to progressively fill the hosts, cannot save costs due to powering off unused nodes because VMs are equitable distributed among physical resources. Therefore, the provider is not able to maximize its profit as much as it does with the mitigation strategy *Reduce* 1.

Under-provisioning. Figs. 7 and 8 illustrate the effects of under-provisioning a private Cloud on the PaaS's weekly profit and its customer satisfaction, respectively. The results are presented for each different risk mitigation strategy (those described in Section 5.3), and depending on the utilization of the private Cloud (i.e. the number of services hosted). As shown in these figures, the most beneficial response in terms of fulfilling both BLOs is to *Share* the risk. In this case, the PaaS outsources the necessary resources to support the operation of services that do not fit into in-house resources. Although the cost of outsourced resources is higher than the cost of local ones, it is also lower than the loss due to SLA violations if deployed locally. Therefore, outsourcing operations are very suitable in order to maximize the profit earned by the provider when its private Cloud is under-provisioned. Furthermore, the compliance with SLAs of services running on those remote resources is ensured by the SLA agreed between the PaaS and third-party IaaS. As a result, this adoption response ensures the potential positive impacts of under-provisioning a private Cloud while, at the same time, avoiding the realization of negative consequences. The *Retain* response is critical for the achievement of both BLOs. This response unconsciously overlooks in-house physical resources, since the private Cloud capacity is not enough to support the load received by the provider. As shown in Fig. 7, the weekly profit is lowered due to large amount of losses incurred by violating SLAs. Finally, the *Avoid* response, which cancels the deployment and operation of the services over the provider's capacity, has the main advantage of ensuring the maximization of customer satisfaction. Note that the profit is always the same since the amount of services running in the private Cloud is the maximum.

6.2.2. Risks of resource failures

As seen in Figs. 9 and 10, which show the effects of resource failures in a private Cloud on the PaaS's weekly profit and its

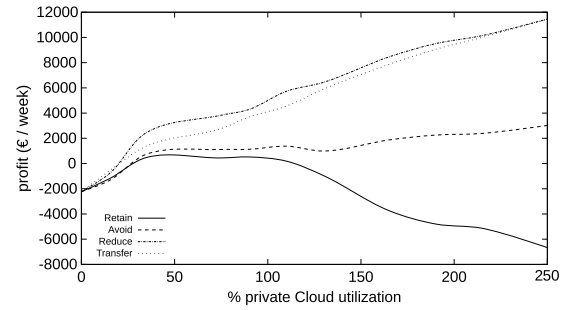


Fig. 9. The effects of resource failures on a provider's weekly profit, depending on both the risk treatment response used (described in Section 5.4) and the number of services hosted.

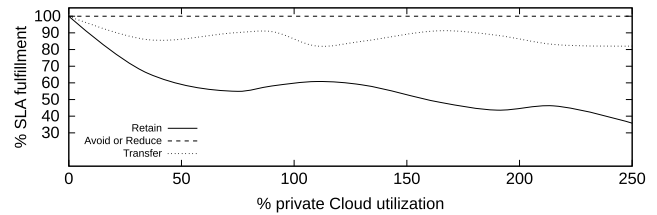


Fig. 10. The effects of resource failures on customer satisfaction, expressed as a percentage of SLA fulfillment, depending on both the risk treatment response used (described in Section 5.4) and the number of services hosted.

customer satisfaction, respectively, the best response to address this threatening risk is to *Reduce* it, by recovering the operation of services upon resource failures. This can be done easily under virtualized environments, by using a checkpointing/restoring mechanism of VMs.

The *Transfer* response, which outsources Gold QoS services, achieves a profit maximization very similar to the best one. This occurs because it ensures the compliance of SLAs with Gold QoS, which generates more benefit for the provider. However, some SLAs with Silver or Bronze QoS categories are violated, since their associated services are running in the faulty private Cloud infrastructure. As shown in Fig. 10, this has an impact on customer satisfaction. The *Avoid* response replicates the amount of virtual resources supporting services' operation and, for this reason, the PaaS also experiences a 100% SLAs fulfillment, but the cost of maintaining the replication of the Cloud infrastructure is too high. This lowers the PaaS's profit.

Finally, the PaaS is subjected to a severe unfulfillment of *ProfMax* and *SatMax* if it decides to retain such risk. Actually, the negative consequences on the whole provider are unexpected if no recovery plan is applied. Notice the profit losses in Fig. 9 due to SLA violations caused by (1) failures of physical and virtual resources and (2) having an under-provisioned infrastructure due to these failures (for 100% of utilization and higher).

6.2.3. Strategy for treatment of Cloud infrastructure-level risks

From the results shown above, we can outline a promising risk treatment strategy to deal with Cloud infrastructure-related risks. It is intended to be used by self-managed Cloud providers aimed at maximizing their profit and customer satisfaction. Such a strategy is composed by the most successful risk responses to deal with each risk evaluated in this section. First, *Reduce* the risk of over-provisioning the private Cloud, by both consolidating services' VMs and aligning the number of online nodes with the time-varying demand received. In addition, services should be highly overbooked (i.e. with an overbooking factor of 1.5) if the utilization of the private Cloud is lower than 15%. Second, *Share* the risk of under-provisioning the private Cloud, by outsourcing the services' operation to public Clouds maintained by third-party IaaS.

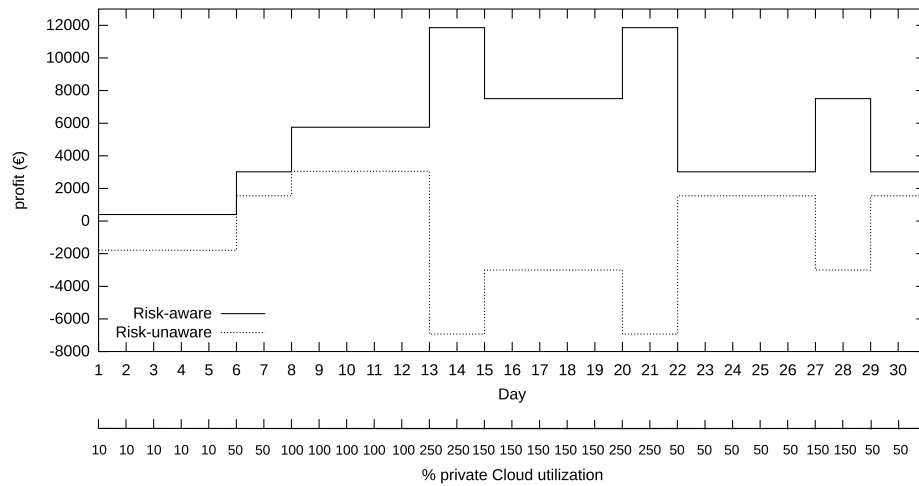


Fig. 11. The effects of self-managing infrastructure-level risks on a provider's profit, depending on the number of services hosted (utilization of the private Cloud).

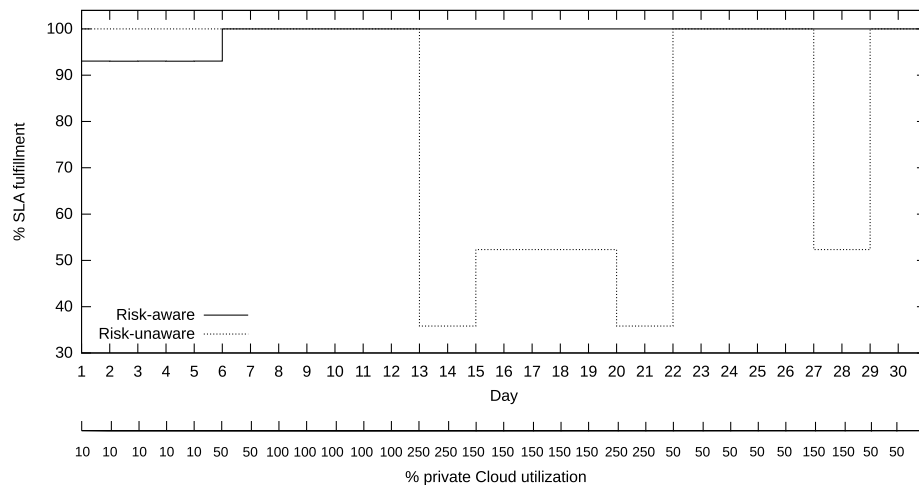


Fig. 12. The effects of self-managing infrastructure-level risks on customer satisfaction, depending on the number of services hosted (utilization of the private Cloud).

Finally, *Reduce* the risk of physical and virtual resource failures, by means of a VM checkpointing/restoring mechanism, for which the execution of services can be recovered upon resource failures.

Figs. 11 and 12 show the effect of such risk treatment strategy on a provider's profit and customer satisfaction, respectively. We compare this strategy with another one that it is unaware of infrastructure-level risks. We experimented with different incoming customer demands (number of hosted services) in order to show the fulfillment of both BLOs in any case, and during a month.

The results demonstrate how a risk-aware policy constantly selects and applies the best risk treatment response(s) by considering the current utilization of the private Cloud. As shown along the x-axis of both figures, the PaaS starts the month with a low demand, which allows it to turn off several of its in-house physical nodes, as well as consolidating and overbooking services over a minimum number of online nodes. These management actions allow the provider to maximize the profit earned, despite the fact of having to pay compensation due to violating roughly 8% of SLAs (if the customer satisfaction BLO is more important for the provider than profit maximization, it could choose a different risk strategy to favor this BLO). Then, it experiences an increase in demand, leading it to completely fill its private Cloud. Note that, in this situation, in which the private Cloud is perfectly provisioned according to users' demands, the differences in BLO fulfillment between both policies analyzed is due to the failures of

resources. In fact, the negative impact of such failures is present throughout the entire month in the case of being unaware of this fact. Later on, the PaaS receives an extra demand which cannot be dealt with by using only the private Cloud resources. For this reason, the risk adoption response of sharing the risk of under-provisioning, which appears at this moment in time, is applied. With this response, the PaaS outsources the operation of services that do not fit into the private Cloud resources to third-party IaaS providers. This risk response assures the SLA fulfillment of all services, either running in in-house or external resources, thus leading to maximizing the profit earned and customer satisfaction. In this situation, a risk-unaware provider is subjected to paying a lot of penalties to its customers since the number of SLA violations is very high. Afterward, the incoming demand drops considerably and, therefore, the private Cloud is over-provisioned again. For this reason, the risk-aware policy decides to turn off half of nodes in order to align the time-varying demand with the number of online nodes. Finally, the PaaS receives a two-day peak in demand which is addressed again by means of outsourcing deals with external IaaS providers.

7. Conclusions

In this paper, we have introduced risk management into the Cloud computing paradigm. In this direction, we first exposed the most important Cloud-specific risks and those for which

Cloud stakeholders are exposed. Afterward, we have presented a Cloud-specific risk management procedure, which is oriented to determine the risk, i.e. the uncertain effect of events (either positive or negative) on the Cloud organization's BLOs. Our risk management process includes the SEmi-quantitative BLO-oriented Cloud Risk Assessment (SEBCRA) procedure, whose main goal is to constantly identify, analyze, and evaluate Cloud risks. Its outcome is a prioritization of risks according to their potential impact(s) on different BLOs. In general, it introduces a methodology that brings transparency to decision-making processes assisted by the risk factor, as well as allowing them to solve uncertainties when self-managing a Cloud provider toward well-defined business objectives and constraints. Furthermore, we have demonstrated how each risk may have dissimilar repercussions on each BLO considered. In fact, the effects of uncertainties highly depend on both the current provider's status and the effectiveness of the risk treatment responses that are being applied.

Finally, and as a use case, we have incorporated our risk management approach in a PaaS provider in order to improve the achievement of the most typical BLOs for Cloud providers, i.e. the maximization of both their profit and customer satisfaction. After using the SEBCRA method, we have observed an urgent need to face the already identified infrastructure-level risks: under- and over-provisioning the private Cloud, and the risk of physical and virtual resource failures. In this regard, we have proposed and evaluated several risk treatment responses to deal with these risks. The results obtained from the experimentation conducted have allowed us to outline a suitable risk management strategy to be embodied in the PaaS provider's self-operation. This strategy is composed by the most efficient responses to address each of those infrastructure-related risks identified, making the provider able to mitigate and adopt their negative and positive consequences.

7.1. Future work

Although we have shown in this paper that providers working in a Cloud environment can benefit from using a risk management methodology for organizations and how this methodology should be adapted to fit in a Cloud organization, our future work includes integrating this risk management methodology in a framework for the self-management of Cloud providers in a large-scale environment. This requires the implementation of an autonomic risk-aware scheduler, which will be based on business-driven policies, management rules, and heuristics that should constantly assist Cloud providers when aiming to improve their profit and reliability, among many others. Moreover, it will assist policies in charge of selecting the best third-party Cloud providers to create multi-Cloud scenarios. Those policies should consider, in a synergistic way, several parameters such as the provider's BLOs, Cloud service requirements, and external providers' dependability. In addition, we are pondering the inclusion of the ALARP (As Low As Reasonably Practicable) principle [53] into the risk analysis subprocess of the SEBCRA method. Basically, a risk is within the ALARP range if the cost needed to reduce it is greatly disproportionate to the benefit gained.

Another important issue will be the tackling of scenarios in which multiple (and contradictory) BLOs are defined by Cloud organizations. In these cases, several trade-offs appear and, therefore, complex business-driven and risk-aware management policies need to be developed. Finally, in this paper, we have assumed a unified viewpoint representing the "business" entity (e.g. group of business executives), which provides a unified view of the organization's BLOs, the potential risks and their impact and, therefore, a single risk prioritization scheme. In practice, this could be difficult to achieve, so additional mechanisms could be needed to get this unified view or to deal with disparate viewpoints within the business executives.

Acknowledgments

This work is supported by the Ministry of Science and Technology of Spain and the European Union (FEDER funds) under contract TIN2007-60625, by the Generalitat de Catalunya under contract 2009-SGR-980, and by the European Commission under FP7-ICT-2009-5 contract 257115 (OPTIMIS).

References

- [1] R. Buyya, C. Yeo, S. Venugopal, Market-oriented cloud computing: vision, hype, and reality for delivering it services as computing utilities, in: 10th IEEE International Conference on High Performance Computing and Communications, 2008. HPCC'08, Dalian, China, 2008, pp. 5–13.
- [2] Gartner, Inc. analysts, Top 10 Strategic Technologies for 2011, Gartner Report (press release).
- [3] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al. Above the clouds: a Berkeley view of cloud computing, EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28.
- [4] Salesforce CRM, (Website, 2011) <http://www.salesforce.com/crm>.
- [5] Google App Engine, (Website, 2011) <http://code.google.com/appengine>.
- [6] Amazon EC2, (Website, 2011) <http://aws.amazon.com/ec2>.
- [7] A. Spellmann, R. Gimarc, M. Preston, O. Innovations, I. Hyperformix, Leveraging the cloud for green IT: predicting the energy, cost and performance of cloud computing, in: Computer Measurement Group Conference, 2009. CMG'09, Dallas, Texas, USA, 2009, pp. 1–17.
- [8] J.O. Fitó, M. Macías, J. Guitart, Toward business-driven risk management for cloud computing, in: Proceedings of the 6th International Conference on Network and Service Management (CNSM 2010), Niagara Falls, Canada, October 25–29, 2010, pp. 238–241.
- [9] ISO 31000:2009, Risk management—Principles and guidelines http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43170.
- [10] ISO Guide 73:2009, Risk Management Vocabulary http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44651.
- [11] D. Macdonald, Practical Machinery Safety, Newnes, 2004.
- [12] Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing V2.1, <http://www.cloudsecurityalliance.org/csaguide.pdf>.
- [13] ENISA Cloud Computing Risk Assessment, <http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment>.
- [14] J. Heiser, M. Nicolett, Assessing the security risks of cloud computing, Gartner Report Available at: <http://www.gartner.com/DisplayDocument?id=685308>.
- [15] H. Motahari-Nezhad, B. Stephenson, S. Singhal, Outsourcing Business to Cloud Computing Services: Opportunities and Challenges, Report HPL-2009-23, HP Labs.
- [16] T. Betcher, Cloud Computing: Key IT-Related Risks and Mitigation Strategies for Consideration by IT Security Practitioners, Ph.D. thesis, University of Oregon, 2010.
- [17] T. Aven, Risk Analysis and Management. Basic Concepts and Principles, Reliability & Risk Analysis: Theory & Applications 2.
- [18] T. Aven, O. Renn, On Risk Defined as an Event where the Outcome is Uncertain, Journal of Risk Research 12 (1) (2009) 1–11.
- [19] T. Aven, A Semi-quantitative Approach to Risk analysis, as an Alternative to QRAs, Reliability Engineering and System Safety 93 (6) (2008) 790–797.
- [20] A. McNeil, et al., Quantitative Risk Management: Concepts, Techniques, and Tools, in: Princeton Series in Finance, 2005.
- [21] J. Hnilica, Qualitative And Semi-Quantitative Risk Analysis Of A Project, Acta Oeconomica Pragensia 2008 (3) (2008) 62–69.
- [22] C. Xenakis, D. Apostolopoulou, A. Panou, I. Stavarakakis, A qualitative risk analysis for the GPRS technology, in: 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, EUC'08, Shanghai, China, December 17–20, 2008, Vol. 02, 2008, pp. 61–68.
- [23] B. Nikolić, L. Ružić-Dimitrijević, Risk assessment of information technology systems, Issues in Informing Science and Information Technology 6.
- [24] Y. Wang, T. Elhag, A Fuzzy Group Decision Making Approach for Bridge Risk Assessment, Computers & Industrial Engineering 53 (1) (2007) 137–148.
- [25] AssessGrid - Advanced Risk Assessment & Management for Trustable Grids, (Website, 2009) <http://www.assessgrid.eu/>.
- [26] K. Djemame, et al. Introducing risk management into the grid, in: 2nd IEEE International Conference on e-Science, e-Science'06, Amsterdam, Netherlands, December 04–06, 2006, 2006, p. 28.
- [27] K. Djemame, I. Gotextitay, J. Padgett, K. Voss, O. Kao, Risk Management in Grids, Market-Oriented Grid and Utility Computing (2009) 335.
- [28] K. Voss, K. Djemame, I. Gotextitay, J. Padgett, AssessGrid, Economic Issues Underlying Risk Awareness in Grids, in: Lecture Notes in Computer Science, vol. 4685, 2007, p. 170.
- [29] K. Djemame, J. Padgett, I. Gourlay, D. Armstrong, Brokering of risk-aware service level agreements in grids, Concurrency and Computation: Practice & Experience 23 (13) (2011) 1558–1582.
- [30] A. Keller, K. Voss, D. Battre, M. Hovestadt, O. Kao, Quality assurance of Grid service provisioning by risk aware managing of resource failures, in: Risks and Security of Internet and Systems, 2008. CRISIS'08. Third International Conference on, 2008, pp. 149–157.

- [31] K. Voss, Risk-aware migrations for preprocessing SLAs, in: International Conference on Networking and Services (ICNS), Silicon Valley, USA, 2006.
- [32] K. Voss, Enhance Self-managing Grids by Risk Management, in: Networking and Services, 2007. ICNS, Third International Conference on, 2007, pp. 27–27.
- [33] Federation of European Risk Management Associations, (Website, 2008) www.ferma.eu.
- [34] C. Yeo, R. Buyya, Integrated risk analysis for a commercial computing service, in: IEEE International Parallel and Distributed Processing Symposium, IPDPS'07, Long Beach, California, USA, March 26–30, 2007, pp. 1–10.
- [35] D. Irwin, L. Grit, J. Chase, Balancing risk and reward in a market-based task service, in: 13th IEEE International Symposium on High Performance Distributed Computing, HPDC'06, Honolulu, Hawaii, USA, June 04–06, 2006, 2004, pp. 160–169.
- [36] F. Popovici, J. Wilkes, Profitable services in an uncertain world, in: 2005 ACM/IEEE conference on Supercomputing, SC'05, Washington, DC, USA, November 12–18, 2005, 2005, p. 36.
- [37] I. Petri, O.F. Rana, Y. Rezgui, G.C. Silaghi, Risk assessment in service provider communities, in: GECON, 2011, pp. 135–147.
- [38] R. Wideman, Project and program risk management: a guide to managing project risks and opportunities, Project Management Institute, 1992.
- [39] ISO/IEC 27005:2008, Information technology - Security Techniques - Information security risk management http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42107.
- [40] S. Srikantaiah, A. Kansal, F. Zhao, Energy aware consolidation for cloud computing, in: Proceedings of the 2008 conference on Power aware computing and systems, HotPower'08, Berkeley, CA, USA, December 8–10, Berkeley, CA, USA, 2008, pp. 10–10.
- [41] A. Verma, P. Ahuja, A. Neogi, pmapper: power and migration cost aware application placement in virtualized systems, Middleware 2008, Leuven, Belgium, December 1–5 2008, pp. 243–264.
- [42] EMOTIVE Cloud, (Website, 2012) <http://www.emotivecloud.net>.
- [43] F. Julià, J. Roldán, R. Nou, J.O. Fitó, A. Vagué, I. Goiri, J.L. Berral, EEFSim: energy efficiency simulator, Tech. rep., Technical University of Catalonia (UPC, UPC-DAC-RR-2010-19 (2010).
- [44] Europe's energy portal, Website, (2011) <http://www.energy.eu>.
- [45] I. Goiri, F. Julià, R. Nou, J. Berral, J. Guitart, J. Torres, Energy-aware scheduling in virtualized datacenters, in: 12th IEEE International Conference on Cluster Computing, CLUSTER'10, Heraklion, Crete, September 20–24, 2010, pp. 58–67.
- [46] S. Greenberg, E. Mills, B. Tschudi, P. Rumsey, B. Myatt, Best practices for data centers: lessons learned from benchmarking 22 data centers, Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings in Asilomar, CA. ACEEE, August 3 (2006) pp. 76–87.
- [47] Rackspace hosting, Website, (2011) <http://www.rackspace.com/>.
- [48] D. Menascé, V. Almeida, L. Dowdy, L. Dowdy, Performance by Design: Computer Capacity Planning by Example, Prentice Hall, 2004.
- [49] SPECweb2009 E-commerce workload, Website, (2011) <http://www.spec.org/web2009/docs/design/EcommerceDesign.html>.
- [50] Amazon EC2 SLA, (Website, 2011) <http://aws.amazon.com/ec2-sla/>.
- [51] J. Dean, Software engineering advice from building large-scale distributed systems, Stanford CS295 class lecture. Available at: 2007 <http://research.google.com/people/jeff/stanford-295-talk.pdf>.
- [52] T. Thein, S. Chi, J. Park, Availability modeling and analysis on virtualized clustering with rejuvenation, International Journal of Computer Science and Network Security (IJCSNS) 8 (9) (2008) 72.
- [53] R. Melchers, On the ALARP Approach to Risk Management, Reliability Engineering and System Safety 71 (2) (2001) 201–208.



J. Oriol Fitó is a Ph.D. student in Computer Science at the Computer Architecture Department of the Technical University of Catalonia (UPC). He obtained his bachelor degree in computer science and his MS degree in the Computer Architectures, Networks and Systems (CANS) Master program at the UPC in 2007 and 2009, respectively. He is currently working at Barcelona Supercomputing Center (BSC) as a research engineer.



Jordi Guitart received his MS and Ph.D. degrees in Computer Science at the Technical University of Catalonia (UPC), in 1999 and 2005, respectively. Currently, he is a collaborator professor in the Computer Architecture Department of the UPC and an associate researcher at Barcelona Supercomputing Center (BSC). His research interests are oriented towards innovative resource management approaches for modern distributed computing systems. He is involved in a number of European projects.