Chapter 1

Meeting Deadlines, Elastically

Einar Broch Johnsen
University of Oslo

Ka I Pun
University of Oslo

Martin Steen
University of Oslo

S. Lizeth Tapia Tarifa
University of Oslo

Ingrid Chieh Yu
University of Oslo

Abstract.
Cloud computing offers a pay-on-demand scalable infrastructure for data processing. Resource-aware services can exploit this infrastructure to elastically adapt to client traffic according to internal resource policies which balance provided QoS with the accrued costs of deployment. This paper presents initial work on worst-case response time analysis for services which distribute tasks to virtual machine instances with different processing speed. We extend JML-like interfaces with response time annotations and develop a Hoare-style proof system to reason about response time guarantees for services expressed in a simple object-oriented language in which dynamically created objects differ in processing capacity. The simplified setting considered in this paper does not consider loops, concurrency, or reflection; we briefly discuss how these restrictions could be lifted.
1.1 Introduction

A cloud consists of virtual computers that are accessed remotely for data storage and processing. The cloud is emerging as an economically interesting model for enterprises of all sizes, due to an undeniable added value and compelling business drivers [12]. One such driver is *elasticity*: businesses pay for computing resources when needed, instead of provisioning in advance with huge upfront investments. New resources such as processing power or memory can be added to a virtual computer on the fly, or an additional virtual computer can be provided to the client application. Going beyond shared storage, the main potential in cloud computing lies in its scalable virtualized framework for data processing. If a service uses cloud-based processing, its capacity can be automatically adjusted when new users arrive or depending on the input size and required response time of different jobs. Another driver is *agility*: new services can be deployed on the market quickly and flexibly at limited cost. This allows a service to handle its users in a flexible manner without requiring initial investments in hardware before the service can be launched.

Today, software is often designed while completely ignoring deployment or based on very specific assumptions, e.g., the size of data structures, the amount of random access memory, and the number of processors. For the software developer, cloud computing brings new challenges and opportunities [23]:

- **Empowering the Designer.** The elasticity of software executed in the cloud gives designers far reaching control over the execution environment’s resource parameters, e.g., the number and kind of processors, the amount of memory and storage capacity, and the bandwidth. In principle, these parameters can even be adjusted at runtime. The owner of a cloud service can not only deploy and run software, but also control trade-offs between the incurred cost and the delivered quality-of-service.

- **Deployment Aspects at Design Time.** The impact of cloud computing on software design goes beyond scalability. Deployment decisions are traditionally made at the end of a software development process: the developers first design the functionality of a service, then the required resources are determined, and finally a service level agreement regulates the provisioning of these resources. In cloud computing, this can have severe consequences: a program which does not scale usually requires extensive design changes when scalability was not considered a priori.

To realize cloud computing’s potential, software must be *designed for scalability*. This leads to a new *software engineering challenge*: how can the validation of deployment decisions be pushed up to the modeling phase of the software development chain without convoluting the design with deployment details?

The EU project Envisage addresses this challenge by extending a design
Meeting Deadlines, Elastically

by contract approach to service-level agreements for resource-aware virtualized services. The functionality is represented in a client layer. A provisioning layer makes resources available to the client layer and determines how much memory, processing power, and bandwidth can be used. A service level agreement (SLA) is a legal document that clarifies what resources the provisioning layer should make available to the client service, what they will cost, and the penalties for breach of agreement. A typical SLA covers two different aspects: (i) the mutual legal obligations and consequences in case of a breach of contract, which we call the legal contract; (ii) the technical parameters and cost figures of the offered services, which we call the service contract.

This paper discusses some initial ideas about applying program verification techniques to models of virtualized services. We consider response time aspects of service contracts and extend JML-like interfaces with response time annotations. This is formalized using \( \mu \text{ABS} \); \( \mu \text{ABS} \) is a restricted version of ABS \cite{27}, an executable object-oriented modeling language developed in the Envisage project to specify resource-aware virtualized services \cite{5,29,30}. Whereas ABS is based on concurrent objects and asynchronous method calls, the work discussed in this paper is restricted to sequential computation and synchronous method calls. In future work, we plan to alleviate these restrictions.

**Paper organization.** Section 1.2 introduces service interfaces with response-time annotations; Sect. 1.3 introduces the syntax of \( \mu \text{ABS} \), the modeling language considered in this paper; Sect. 1.4 demonstrates the approach on an example; Sect. 1.5 develops a Hoare-style proof system for \( \mu \text{ABS} \); Sect. 1.6 discusses related work; and Sect. 1.7 concludes the paper by a discussion of the limitations and possible extensions of the current work.

### 1.2 Service Contracts as Interfaces

Service level agreements express non-functional properties of services (service contracts), and their associated penalties (legal contracts). Examples are high water marks (e.g., number of users), system availability, and service response time. Our focus is on service contract aspects of client-level SLAs, and on how these can be integrated in models of virtualized services. Such an integration would enable a formal understanding of service contracts and of their relationship to the performance metrics and configuration parameters of the deployed services. Today, client-level SLAs do not allow the potential resource usage of a service to be determined or adapted when unforeseen changes to resources occur. This is because user-level SLAs are not explicitly related to actual performance metrics and configuration parameters of the services; for example, the user-level SLAs may be concerned with end-user response times but not about the number of virtual machine instances on which the service is deployed. The integration of service contracts and configuration parameters...
From Action System to Distributed Systems: The Refinement Approach

```
type Photo = Rat; // size of the file

interface PhotoService {
  @requires \forall p:Photo . p ∈ film && p < 4000;
  @ensures reply == True;
  @within 4 * length(film) + 10;
  Bool request(List<Photo> film);
}
```

**FIGURE 1.1:** A photo printing shop in µABS.

in service models enables the design of resource-aware services which embody application-specific resource management strategies [23].

The term **design by contract** was coined by Bertrand Meyer referring to the contractual obligations that arise when objects invoke methods [36]: only if a caller can ensure that certain behavioral conditions hold before the method is activated (the precondition), it is ensured that the method results in a specified state when it completes (the postcondition). Design by contract enables software to be organized as encapsulated services with interfaces specifying the contract between the service and its clients. Clients can “program to interfaces”; they can use a service without knowing its implementation. We aim at a design by contract methodology for SLA-aware virtualized services, which incorporates SLA requirements in the interfaces at the application-level to ensure the QoS expectations of clients.

We consider an object-oriented setting with service-level interfaces given in a style akin to JML [11] and Fresco [49]; requires- and ensures-clauses express each method’s functional pre- and postconditions. In addition, a **response time guarantee** is expressed in a **within**-clause associated with the method. The specification of methods in interfaces is illustrated in Figure 1.1.

### 1.3 A Kernel Language for Virtualized Computing

ABS supports modeling the deployment of objects on virtual machines with different processing capacities [5, 27, 30]. The µABS language simplifies ABS by letting each object have a dedicated processor with a given processing capacity. Thus, objects are dynamically created instances of classes such that the resource capacity of each object reflects the provisioning contract between that object and its resource provider. In contrast to ABS, communication between named objects is synchronous, which means that a method call blocks the caller until execution has finished. For simplicity in this paper, the objects share a thread of execution where at most one task is **active** and the others are waiting to be executed on the task stack. Although µABS is currently
restricted to sequential execution, execution is elastic in the sense that several objects may provide instances of the same service at different speeds (and different associated costs for the service provider) and the choice of service instance for a task can be dynamically decided by the service based on, e.g., the deadline of the task and the accumulated cost of running the service.

μABS is strongly typed: for well-typed programs, invoked methods are understood by the called object. μABS includes the types Capacity, Cost, and Duration which all extend Rat with an element infinite: Capacity captures the processing capacity of virtual machines per time interval, Cost the processing cost of executions, and Duration time intervals.

Figure 1.2 presents the syntax of μABS. A program $P$ consists of interface and class definitions, and a main block $\{T; x; sr\}$. Interfaces $IF$ have a name $I$ and method signatures $Sg$. Classes $CL$ have a name $C$, optional formal parameters $T x$, and methods $M$. A method signature $Sg$ has a list of specifications $Spec$, a return type $T$, a method name $m$, and formal parameters $x$ of types $T$. In specifications (see Sect. 1.2), assertions $\phi$ express properties of local variables in an assertion language extending the expressions $e$ with logical variables and operators in a standard way; a reserved variable reply captures the method’s return value. A method $M$ has a signature $Sg$, a list of local variable declarations $x$ of types $T$, and statements $sr$. Statements may access local variables and the formal parameters of the class and the method.

Statements are standard, except $job(e)$ which captures an execution requiring $e$ processing cycles. A job abstracts from actual computations but may depend on state variables. Right-hand sides $rhs$ include expressions $e$, object creation $new\ C(\tau)$ with $e$ and synchronous method calls $e.m(\tau)$. Objects are created with a given capacity, which expresses the processing cycles available to the object per time interval when executing its methods. Thus, different instances may have different capacities (and consequently, their usage may

\[ P ::= \begin{array}{l} \{T; C; I; m\} \{T; x; sr\} \\
T ::= C \mid I \mid \text{Capacity} \mid \text{Cost} \mid \text{Duration} \mid \text{Bool} \mid \text{Rat} \\
IF ::= \text{interface } I \{ Sg \} \\
Sg ::= \text{Spec}; m \{T x\} \\
Spec ::= \text{requires } \phi x | \text{ensures } \phi x | \text{within } \phi x \\
CL ::= \text{class } C(T x) \{ M \} \\
M ::= Sg \{T x; sr\} \\
sr ::= s; s \mid x = rhs | \text{job}(e) | \text{if } e \{ s \} \text{ else } \{ s \} \\
rhs ::= e | new\ C(\tau) \text{ with } e | e.m(\tau) \\
e ::= \text{this capacity; deadline } x | v | e \text{ op } e \end{array} \]
have different costs in the metered setting of elastic computing). Method calls in µABS are blocking. Expressions $e$ include operations over declared variables $x$ and values $v$. Among values, $b$ has type $\text{Bool}$, $i$ has type $\text{Rat}$ (e.g., $5/7$), $k$ has type $\text{Capacity}$, $c$ has type $\text{Cost}$, and $d$ has type $\text{Duration}$. Among binary operators $\text{op}$ on expressions, note that division $c/k$ has type $\text{Duration}$. Expressions also includes the following reserved read-only variables: this refers to the object identifier, capacity refers to the processing speed (amount of resources per time interval) of the object, and deadline refers to the local deadline of the current method. (We assume that all programs are well-typed and include further functional expressions and data types when needed in the example.)

Time. µABS has a dense time model, captured by the type $\text{Duration}$. The language is not based on a (global) clock, instead each method activation has an associated local counter deadline, which decreases when time passes. Time passes when a statement $\text{job}(e)$ is executed on top of the task stack. The effect of executing this statement on an object with capacity $k$, is that the local deadline of every task on the stack decreases by $c/k$, where $c$ is the value resulting from evaluating $e$. The initial value of the deadline counter stems from the service contract; thus, a local counter which becomes negative represents a breach of the local service contract. For brevity, we do not present the formal semantics.

### 1.4 Example: A Photo Printing Shop

Let us consider a photo shop service which retouches and prints photos. It is cheaper for the photo shop service to retouch and print photos locally, but it can only deal with low resolution photos in time. For larger photos, the photo shop service relies on using a faster and more expensive laboratory in order to guarantee that all processing deadlines are met successfully.

In this example, a film is represented as a list of photos and, for simplicity, a photo by the size of the corresponding file. As shown in the class diagram of Figure 1.3, an interface PhotoService provides a single method request which handles customer requests to the photo shop service. The interface is implemented by a class PhotoServiceImpl, which has methods retouch for retouching and print for printing a photo, in addition to the request method of the interface. For faster processing, two interfaces FastEdit and FastPrint, which also provide the methods retouch and print, may be used by PhotoServiceImpl. The sequence diagram in Figure 1.4 shows how a photo is first retouched, then printed. The services of retouching and printing are done locally if possible, otherwise they are forwarded to and executed by objects with higher capacities.

The µABS model of the example (Figure 1.5) follows the design by contract approach and provides a contract for every method declaration in an interface
Meeting Deadlines, Elastically

and method definition in a class. These specifications are intended to guarantee that a request to a PhotoService object will not break the specified contract. If we consider the contract for request in more detail, we see that the response time of a request(film) call depends on the length of the film and assumes that the size of every photo contained in the film is smaller than 4000. The implementation of the request method is as follows: Take the first photo in the film (by applying the function head(film)) and check if this photo is low resolution compared to the capacity of the PhotoService object, represented by a size smaller than 500 and a capacity of at least 100, respectively. In this case, the retouch can be done locally, otherwise retouch is done by an auxiliary FastEdit object. A similar procedure applies to printing the retouched photos. Thus, photos of small sizes are retouched and printed locally, while photos with bigger sizes are sent to be retouched and printed externally. The ability to send tasks to the laboratory which meets the deadline at the lowest cost, expresses elasticity in the setting of this example. The implementations of the different methods are abstractly captured using job statements. The expressions ε inside the job statements, could be further refined or calibrated using SACO [3].

FIGURE 1.3: A class diagram for a photo printing shop
From Action System to Distributed Systems: The Refinement Approach

FIGURE 1.4: A sequence diagram for a photo printing shop

1.5 Proof System

Virtual machines are subject to failures. We interpret non-termination of a job statement as an underlying failure and restrict our analysis to partial correctness. The proof system for μABS is formalized as Hoare triples \([\phi] s \{\psi\}\) with a standard partial correctness semantics: if the execution of \(s\) starts in a state satisfying the precondition \(\phi\) and the execution terminates, the result will be a state satisfying the postcondition \(\psi\). In this paper, we are particularly interested in assertions about the deadline variables of method activations.

The reasoning rules for μABS are presented in Figure 1.6. Reasoning about sequential composition, conditional, and assignment statements is standard, and captured by the rules Comp, Cond, and Assign, respectively. Time only passes when \(\text{job}(e)\) is executed; \(\text{job}(e)\) has a duration \(e/cap\) when executed on an object with capacity \(cap\). The assertion in Rule Job ensures that this duration is included in the response time after executing \(\text{job}(e)\). The subsumption rule allows to strengthen the precondition and weaken the postcondition. For method definitions, the premise of Rule Method assumes that the execution of \(sr\) starts in a state where the requires-clause \(\phi\) is satisfied and that the
Meeting Deadlines, Elastically

FIGURE 1.5: A photo printing shop in μABS

expected response time (deadline) is larger than expression e, where e is the specified response time guarantee from the within-clause. When the execution of sr terminates, the result will satisfy the ensures-clause ψ and the expected response time remains non-negative. For method invocations in Rule CALL, the specification of the method is updated by substituting the formal parameters $f^m_p$ by the input expressions $p$. The logical variables for the return value of the method (reply) and of the expected response time are renamed with fresh variables $\alpha$ and $\beta$, respectively. To avoid name clashes between scopes, we assume renaming of other variables as necessary. Object creation (in Rule NEW) is handled similarly to assignment. The precondition ensures that the newly created object of a class $C$ with capacity $e$ correctly implements interface $T$, where $T$ is the type of $x$. (Note that the class instance may or may not implement an interface, depending on its capacity.) If a method has a return
value, expression $e$ in the return statement will be assigned to the logical variable $\text{reply}$ in Rule Return, and can be handled by the standard assignment axiom in Rule Assign.

Although $\mu$ABS does not currently include loops, the language supports recursion through interfaces with associated contracts. This suggests how loops can be handled in the proof system, in terms of loop invariants which express execution time for the remaining iterations of the loop.

Example. We show in Equation 1.3 the skeleton of the proof for the method request in Figure 1.5 by using the proof system presented in Figure 1.6. Let $sr$ refer to the method body of request and let $s$ denote $sr$ without the return statement. In addition, we introduce the following abbreviations for formulas:

\[
\psi = \text{reply} == \text{True}, \\
\psi_1 = \psi \land \text{deadline} \geq 0, \\
\phi = \forall p : \text{Photo}, p \in \text{film} \land p < 4000, \text{ and} \\
e = 4 \ast \text{length} (\text{film}) + 10
\]

and assume that

\[
\psi_2 = \text{reply} == \text{deadline} \geq 0 \land \text{deadline} \geq 0
\]
Meeting Deadlines, Elastically

is the postcondition of the assignment \( \text{reply} = \text{deadline} \geq 0 \).

By Rule \textsc{Method}, the assertions \( \phi \) and \( \text{deadline} > e \) serve as precondition to the whole method body \( sr \), where \( \phi \) and \( e \) are given in the \textsc{requires}- and \textsc{within}-clauses associated with the definition of the method \text{request} in Figure 1.5. The postcondition of the method body consists of \( \psi \), which is specified in the \textsc{ensures}-clause as \( \text{reply} = \text{True} \), and the expression \( \text{deadline} \geq 0 \). Rule \textsc{Return} converts the return-statement into a statement where the expression \( \text{deadline} \geq 0 \) is assigned to the logical variable \( \text{reply} \). Then, by the assignment axiom \textsc{Assign}, and with the postcondition \( \psi_2 \) assumed in Equation 1.2, the precondition \( \psi_3 \) is the postcondition with the logical variable \( \text{reply} \) substituted with the expression \( \text{deadline} \geq 0 \), and thus \( \psi_3 = \text{True} \land \text{deadline} \geq 0 \). By using Rule \textsc{Subsumption}, the postcondition \( \psi_2 \) is weakened to the given postcondition \( \psi_1 \). By Rule \textsc{Comp}, the assertion \( \psi_3 \) is also the postcondition of the statement \( s \).

\[
\begin{align*}
\{ \phi \land \text{deadline} > e \} & \ s \ \{ \psi_3 \} & \{ \psi_3 \} & \text{reply} = \text{deadline} \geq 0 & \{ \psi_2 \} & \psi_2 \Rightarrow \psi_1 \\
\{ \phi \land \text{deadline} > e \} & \ s \ \{ \psi_3 \} & \{ \psi_3 \} & \text{reply} = \text{deadline} \geq 0 & \{ \psi_1 \} \\
\end{align*}
\]

\textbf{requires} \( \phi \); \textbf{ensures} \( \psi \); \textbf{within} \( e \);

\texttt{Bool request(List(Photo) film){sr}}

(1.3)

For brevity, the rest of the proof is omitted. The proof can be completed by repeatedly applying the corresponding rules from the presented proof system.

1.6 Related Work

The work presented in this paper is related to the ABS modeling language and its extension to virtualized computing on the cloud, developed in the Envisage project [4]. ABS [27] and its extensions with time [10], deployment component and resource-awareness [30] provide a formal basis for modeling virtualized computing. ABS has been used in two larger case studies addressing resource management in the cloud by combining simulation techniques and cost analysis, but not by means of deductive verification techniques; a model of the Montage case study [14] is presented in [29] and compared to results from specialized simulation tools and a large ABS model of the Fredhopper Replication Server has been calibrated using SACO [3] (a cost analysis tool for ABS) and compared to measurements on the deployed system in [5, 13]. The
main difference between resource-aware models in ABS [30] and the work presented in this paper, is an elimination of non-determinism; in addition to the restriction to sequential programs discussed above, \( \mu \)ABS unifies objects and deployment components, such that objects cannot compete for the resources on a server. As we extend \( \mu \)ABS towards ABS in future work, this seems like a reasonable restriction to enable more precise reasoning, but it could mean that object creation could fail if the targeted deployment component lacks resources! Related techniques for modeling deployment in embedded real-time systems may be found in an extension of VDM++ [48]. In this extension, static architectures are explicitly modeled using CPUs and buses. The approach uses fixed resources targeting the embedded domain. Whereas ABS has been designed to support compositional verification based on traces [15], neither ABS nor VDM++ supports deductive verification of non-functional properties today.

Assertional proof systems addressing timed properties, and in particular upper bounds on execution times of systems, have been developed, the earliest example perhaps being [44]. Another early example of reasoning about real-time is Nielson’s extension of classical Hoare-style verification to timed properties of a given program’s execution [39, 40]. Soundness and (relative) completeness of the proof rules of a simple while-language are shown. Shaw [43] presents Hoare logic rules to reason about the passage of time, in particular to obtain upper and lower bounds on the execution times of sequential, but also of concurrent programs.

Hooman’s work on assertional reasoning and Hoare logic [25] for concurrent programs covers different communication and synchronization patterns, including shared-variable concurrency and message passing using asynchronous channels. The logic introduces a dense time domain (i.e., the non-negative reals, including \( \infty \)) and conceptually assumes a single, global clock for the purpose of reasoning. The proof system is developed for a small calculus focusing on time and concurrency, where a delay-statement can be used to let time pass. This is comparable to the job-expression in our paper, but directly associates a duration with the job. In contrast, we associate a cost with the job, and the duration depends on the execution capacity of the deployed object where the statement is executed. Timed reasoning using Dijkstra’s weakest precondition formulation of Hoare logic can be found in [21]. Lamport’s temporal logic of actions TLA [1, 35] has likewise been extended with the ability to reason about time [34]. Similar to the presentation here, the logical systems are generally given by a set of derivation rules in a pre-/post-condition style. Similar to our work, these approaches are compositional in that timing information for composed programs, including procedure calls, is derived from that of more basic statements. While being structural in allowing syntax-directed reasoning, these formalisms do not explore timed interfaces as part of the programming calculus as we have done here. Thus, these approaches do not support the notion of design-by-contract compositionality for non-functional properties that has been suggested in this paper.
Complementing the theoretical development of proof systems for real-time properties, corresponding reasoning support has been implemented within theorem provers and proof-assistants, for instance for PVS in [17] (using the duration calculus), and HOL [20]. An interesting approach to compositional reasoning about timed system is developed in [18]. As its logical foundation, the methodology uses TRIO [19], a general-purpose specification language based on first-order linear temporal logic. In addition, TRIO supports object-oriented structuring mechanisms such as classes and interfaces, inheritance, and encapsulation. To reason about open systems, i.e., to support modular or compositional reasoning, the methodology is based on a rely/guarantee formalization and corresponding proof rules are implemented within PVS. Similarly, a rely/guarantee approach for compositional verification in linear-time temporal logics is developed in [31,47]. A further compositional approach for the verification of real-time systems is reported in [26], but without making use of a rely/guarantee framework.

Refinement-based frameworks constitute another successful design methodology for complex system, orthogonal to compositional approaches. Aiming at a correct-by-construction methodology, their formal underpinning often rests on various refinement calculi [7, 37, 38]. Refinement-based frameworks have also been developed for timed systems. In particular, Kaisa Sere and her co-authors [9] extended the well-known formal modeling, verification, and refinement framework Event-B [2] with a notion of time, resulting in a formal transformational design approach where the proof-obligations resulting from the timing part in the refinement steps are captured by timed automata and verified by the Uppaal tool [8].

The Java modeling language JML [11] is an interface specification language for Java which was used as the basis for the interface specification of service contracts in our paper. Extensions of JML have been proposed to capture timed properties and to support component-based reasoning about temporal properties [32, 33]. These extensions have been used to modularly verify so-called performance correctness [45,46]). For this purpose, JML’s interface specification language is extended with a special duration-clause, to express timing constraints. The JML-based treatment of time is abstract insofar as it formalizes the temporal behavior of programs in terms of abstract “JVM cycles”. Targeting specifically safety critical systems programmed in SCJ (Safety-critical Java), SafeJML [22] re-interprets the duration-clause to mean the worst-case execution time of methods concretely in terms of absolute time units. For a specific hardware implementation for the JVM for real-time applications, [42] presents a different WCET analysis [41] for Java. The approach does not use full-fledged logical reasoning or theorem proving, but is a static analysis based on integer linear programming and works at the byte-code level. We are not aware of work relating real-time proof systems to virtualized software, as addressed in this paper.
1.7 Discussion

Cloud computing provides an elastic but metered execution environment for virtualized services. Services pay for the resources they lease on the cloud, and new resources can be dynamically added as required to offer the service to a varying number of end users at an appropriate service quality. In order to make use of the elasticity of the cloud, the services need to be scalable. A service which does not scale well may require a complete redesign of its business code. A virtualized service is able to adapt to the elasticity provided by the cloud. We believe that the deployment strategy of virtualized services and the assessment of their scalability should form an integral part of the service design phase, and not be assessed a posteriori after the development of the business code as it is done today. The design of virtualized services provides new challenges for software engineering and formal methods.

Virtualization empowers the designer by providing far-reaching control over the resource parameters of the execution environment. By incorporating a resource management strategy which fully exploits the elasticity of the cloud into the service, resource-aware virtualized services are able to balance the service contracts that they offer to their end users, to the metered cost of deploying the services. For resource-aware virtualized services, the integration of resource management policies in the design of the service at an early development stage seems even more important.

This paper pursues a line of research addressing the formal verification of service contracts for virtualized services. We have considered a very simple setting with an interface language which specifies services, including their service contracts in the form of response time guarantees, and a simple object-oriented language for realizing these services. To support non-functional behavior, the language is based on a real-time semantics and associates deadlines with method calls. Virtualization is captured by the fact that objects are dynamically created with associated execution capacities. Thus, the time required to execute a method activation depends not only on the actual parameters to the method call, but also on the execution capacity of the called object. This execution capacity reflects the processing power of virtual machine instances, which are created from within the service itself. The objective of the proof system proposed in this paper is to apply deductive verification techniques to ensure that all local deadlines are met during the execution of a virtualized service. This proof system builds on previous work for real-time systems, and recasts the deductive verification of timing properties to a setting of virtualized programs. The extension of service interfaces with response-time guarantees, as proposed in this paper, allows a compositional design-by-contract approach to service contracts for virtualized systems.

Whereas our work goes in the direction of worst-case cost analysis, it would also be interesting to consider soft real-time requirements as typically en-
countered in service-oriented computing. This could in principle be done by incorporating probabilistic information about response times and execution cost into the models. As such, the approach taken in this paper could be complemented by simulations and statistical techniques (e.g., Monte Carlo simulations have been applied in the context of ABS [28]). However, as cloud computing is increasingly used for critical services in domains such as health and banking, we believe that there is a need for analysis techniques for hard deadlines also in the context of virtualized services.

Several challenges to the proposed approach are left for future work, in particular the extension to concurrency and asynchronous method calls. Currently we plan to address this challenge in terms of proof rules which make use of explicit assumptions about the size of the queues for concurrent ABS objects. The size of the queues can be statically detected; e.g., the size may be approximated by techniques such as may-happen-in-parallel analysis [16]. Furthermore, in the concurrent setting, it is interesting to work with reflection; e.g., an object may query the current load of its virtual machine and use this as a basis for resource management. In this paper, we have considered the explicit allocation of resources for each object. In future work, it would be advantageous to lift the bounded queue length from individual objects to groups of objects, and work with the deployment of such groups. Another challenge is the incorporation of code which reflects the actual computations (replacing the job-statements of this paper). In this case, the abstraction to job-statements could be done by incorporating a worst-case cost analysis [3] into the proof system. Another interesting challenge, which remains to be investigated, is how to incorporate the global requirements which we find in many service-level agreements into a compositional proof system, such as the maximum number of end users.
Bibliography


References


Bibliography


Bibliography


