
Proceedings Paper:
Armstrong, D, Djemame, K, Espling, D et al. (2 more authors) (2013) Runtime virtual machine recontextualization for clouds. In: Caragiannis, I, (ed.) Euro-Par 2012: Parallel Processing Workshops. Revised Selected Papers. Euro-Par 2012, 27-31 Aug 2012, Rhodes Islands, Greece. Lecture notes in computer science . Springer , 567 - 576. ISBN 978-3-642-36948-3

https://doi.org/10.1007/978-3-642-36949-0_66

Runtime Virtual Machine Recontextualization for Clouds

Django Armstrong[1], Daniel Espling[2], Johan Tordsson[2], Karim Djemame[1], and Erik Elmroth[2]

[1] University of Leeds, United Kingdom,
[2] Umeå University, Sweden

Abstract

We extend cloud application contextualization, i.e. the dynamic configuration of a VM upon initialization, to leverage the full potential of cloud hosting by introducing the idea of recontextualization. We present a runtime cloud recontextualization mechanism to allow VM images and instances to be dynamically re-configured without restarts or downtime. The mechanism is applicable to all aspects of configuring a VM from virtual hardware to multi-tier software stacks, without the need to customize the guest VM. We present our work via a use case: the reconfiguration of a cross-cloud migratable monitoring service in a dynamic cloud environment. We discuss the details of the interoperable recontextualization mechanism, its architecture and demonstrate a proof of concept implementation through a performance evaluation. The results of this evaluation show that the solution performs adequately with an overhead of 18% of the total migration time, illustrating the feasibility of the solution.

1 Introduction

Infrastructure as a Service (IaaS) clouds are commonly based on virtualized hardware platforms executing and orchestrating self-contained Virtual Machines (VMs), which are comprised of multiple virtual devices. Several VM instances can be started using the same master disk image and each new VM instance is uniquely configured, *contextualized*, with instance specific settings prior to (or at the early stages of) VM execution.

The life-cycle of a cloud application is comprised of three individual phases as shown in Figure 1. The Construction phase refers to the development of a cloud application making use of platform services and dividing that application into a set of VM images. In the Deployment phase a constructed application is deployed on to suitable infrastructure and finally in the Operation phase the cloud application is executed. The application can be configured with general settings in the Construction phase and contextualized with settings specific to the provider environment in the Deployment phase. Recontextualization offers dynamic reconfiguration of any kind of setting in the Operation phase.

Recent work on IaaS systems have a lot in common with the vision of autonomic computing, as outlined by Kephart and Chess [9]. One of the major

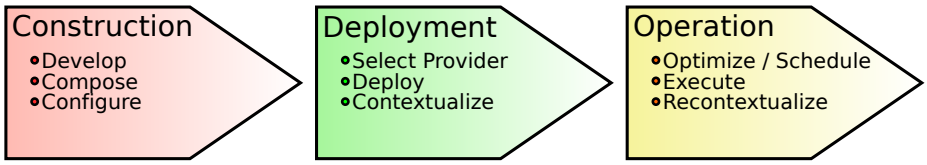

Fig. 1. The life-cycle of a cloud application.

aspects of autonomic computing that has yet to be realized is self-configuration, the automated configuration and adjustment of systems and components. Our earlier work on contextualization [1] presents a mechanism for boot-time self-configuration of VMs. This work extends state of the art and our earlier efforts by introducing runtime recontextualization, enabling adaptation of VM behavior in response to internal changes in the application to which the VM belongs or to external changes affecting the execution environment of the VM.

The contributions of this paper are: i) The development of an architecture and mechanism for the purpose of recontextualization. ii) An implementation and evaluation of a recontextualization system. iii) Practical experiences with hypervisor interoperability. The rest of this paper is organized as follows: Section 2 outlines the problem to be solved, a set of requirements and an illustrative scenario of recontexualization for application monitoring. Section 3 discusses different approaches considered for runtime recontextualization. Section 4 presents our proposed solution for recontextualization of VMs including an evaluation of the approach. Finally, a conclusion and future work are presented in Section 5.

2 Problem Statement and Requirements

A motivational factor behind the need for runtime recontextualization stems from VM migration in clouds [3,14]. Using migration, a VM can be transferred from one physical host to another without explicitly shutting down and subsequently restarting the VM [4]. The entire state of the VM, including e.g., memory pages, are transferred to the new host and the VM can resume its execution from its state prior to migration. As a consequence, no contextualization is triggered again when the VM is resumed.

As presented in [6], there are several different cloud scenarios:

- Bursting - The partial or full migration of an application to a third party IaaS provider, for example when local resources are near exhaustion.
- Federation - The migration of an applications workload between a group of cooperating IaaS providers, e.g., when a single provider's resources are insufficient for application redundancy.
- Brokering - The migration of an application's VMs, e.g., for the purpose of maintaining an agreed QoS when a broker is used to select appropriate IaaS providers.

In all these cloud scenarios VM migration is a necessity, e.g., for the purpose of consolidating resources and maintaining levels of Quality of Service (QoS). These scenarios can be used to define the requirements for any potential recontextualization mechanism. Using them we have identified the following requirements as the most important:

i. A triggering mechanism for recontextualization on VM migration.
ii. A secure process to gather and recreate contextualization data after migration.
iii. A hypervisor agnostic solution that maintains IaaS provider interoperability.
iv. An approach that limits the pervasive nature of the solution and minimizes modifications at the IaaS level.

Each of these scenarios require recontextualization at runtime. In the Bursting scenario, if an IaaS provider is not obligated to divulge third party providers used for outsourcing of computational resources, an application may end up deployed on to a third party's infrastructure that requires the use of their local infrastructure services. A dynamic federation of IaaS providers created during negotiation time that alters during the operation phase requires infrastructure services to be discovered dynamically. The same is applicable in the case of a Broker, knowledge of an IaaS provider local infrastructure services is not available during deployment until after the Broker has selected an appropriate provider.

Specifically, the lack of knowledge on the attributes of an IaaS provider's local infrastructure service available during deployment time, motivates our work. An example of such a service that exhibits configuration issues is application-level monitoring. In the scenarios, the monitoring service endpoint attribute of an application is not available for a contextualization system to use to configure monitoring probes during the deployment phase of a cloud application's life cycle. When an attribute is available there is the possibility that it can change over the operation of the application. These issues motivate the need for a mechanism to fetch configuration data during application operation and provide new context to application dependencies, thus *recontextualization*.

Recontextualization can be used to adapt to any system changes, including making newly migrated VMs operate properly in the (potentially different) system environment of a new host and adapt to different application middleware services at the PaaS level through the dynamic binding of APIs enabling the execution of site specific code, the latter of which is out of scope for this paper. In the following section, we illustrate recontextualization with service-level monitoring [7] as an example scenario.

2.1 Example Scenario

A typical cloud application must be continually monitored during runtime, and a sample configuration for application monitoring is shown in Figure 2. Monitoring data can be used for several purposes, e.g., for automatic application scaling or to assess the likelihood of breaching its Service Level Agreement (SLA). Application level metrics (also called Key Performance Indicators) are sent from inside the VM to an external monitoring endpoint for processing.

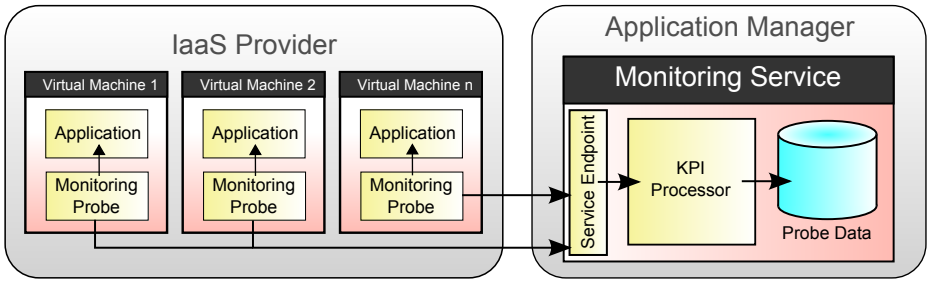

Fig. 2. Monitoring applications in a IaaS provider.

Each monitoring probe that gathers KPI data must be configured with the endpoint of the monitoring service. The endpoint can be associated with a IaaS specific service or a service running at a remote location and depends on what entity within connected clouds has control over application management. When deploying to a IaaS provider, the endpoint for the monitoring service is configured using contextualization in the Deployment phase. However, in a multi-site scenario the VM may be migrated to another site during runtime, and VMs migrating from one site to another must therefore be dynamically recontextualized with this information in the Operation phase.

3 Recontextualization Approaches

As far as we are aware no previous approach supporting recontextualization exists, but instead several different approaches used for contextualization have been considered for use in recontextualization. Keahey and Freeman [8] present fundamental work on contextualization in virtual clusters. Recontextualization is mentioned but deemed out of scope for their work.

The recontextualization procedure has two major obstacles that have to be dealt with by any approach; how is recontextualization triggered and where can the necessary information be found? Below are some alternatives for recontextualization, listed and discussed, from the perspective of the above two challenges.

Contextualized direct addressing is based on a known endpoint address that is specified in the initial contextualization phase, as described by Armstrong et al. in [1]. A similar approach is used for Puppet [13], a mass-machine configuration tool for HPC-like environments. During operations, this endpoint address is queried for the remaining, updated context information. Furthermore, this approach is interoperable and requires no host and hypervisor modifications, but requires that the end point address is constant when the VM is migrated to other domains. This approach offers no procedure for triggering a new phase of recontextualization, and has to rely on periodically querying the endpoint for updated information.

Hypervisor network proxying also relies on periodically querying an external endpoint address for context information, but in this method a standard virtual network address is used and the hypervisor (and associated virtual network management) is responsible for routing this call to a host specific endpoint. This approach, used by Clayman et al. in [5], is transparent to the VM but requires modifications to a hypervisor.

Hypervisor interaction from the guest can be used to offer contextualization straight from the hypervisor itself, using e.g. a customized API both to react to changes in context information and to transfer new information. However, this solution requires modifications both to hypervisor and guest operating system software and would require considerable standardization to be widely available, due to compatibility of virtual hardware APIs between hypervisor technologies.

Dynamic virtual device mounting is based on dynamically mounting virtual media containing newly generated content in a running VM via the reuse of existing hypervisor interfaces and procedures [1]. Interoperability is achieved by reusing existing drivers for removable media such as USB disks or CD-ROM drives. Recontextualization can be detected by the guest OS by reacting to events triggered when new USB or CD-ROM media is available.

We find the dynamic virtual device mounting to be the most promising solution due to the inherent interoperability and support in all major operating system. The ability to manage virtual devices is also offered by the Libvirt API [11], inferring that there is fundamental support for these operations in most major hypervisors. The following section describes our recontextualization solution in more detail.

4 A Recontextualization Solution

In this section, an implementation of a system for runtime recontextualization is described, followed by an evaluation to validate the suggested approach. The idea is to use virtual device mounting techniques in response to migration events and thus enable automatic self-configuration of newly acquired VMs. The following subsections discuss the mechanism, architecture, and evaluation in more detail.

4.1 Mechanism

Figure 3 illustrates the recontextualization approach used in the implementation. Each VM is assigned a virtual CD-ROM device for contextualization on which the host-specific contextualization data can be found. When a VM is migrated from one host to another events are triggered by the hypervisor. In response to these events, the recontextualizer software triggers a detachment of the virtual device mounted with contextualization information, and once the migration is completed a new virtual device with context information relevant for the new host is automatically attached to the VM as it resumes operation after migration.

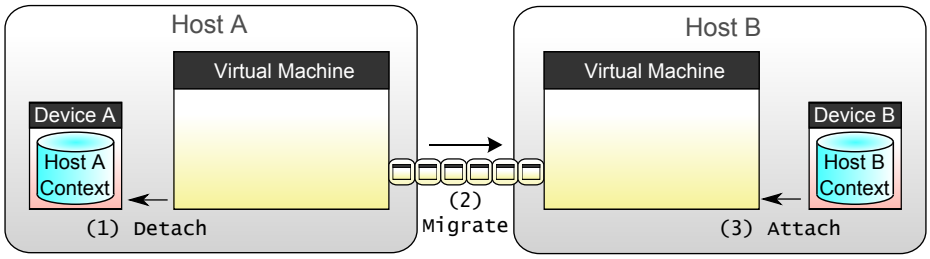

Fig. 3. Recontextualization approach overview.

Event support including migrations is present in several hypervisors, including Xen [2] and KVM [10]. The Libvirt API makes a unified approach to VM management available and includes event support. An initial version of the recontextualization system was implemented using KVM with QEMU [12] specific events and control APIs, and the second version was implemented using Libvirt to make the solution hypervisor independent. Libvirt provides a number of event types that can be monitored via a callback: i) Started, ii) Suspended, iii) Resumed, iv) Stopped, and v) Shutdown. Upon receiving an event callback details are returned on the specific cause of the event, for example the shutting down of VM on a host machine triggered by migration terminating successfully.

4.2 Architecture

The architecture of the implemented system is shown in Figure 4. Up-to-date context data is dynamically bundled as ISO images on the host. The recontextualizer, implemented in Python, manages the attachment and detachment of virtual CD-ROM devices inside a VM that contain the data held within the ISO image in response to events from the hypervisor. The Python Libvirt API bindings are used to access the Libvirtd daemon for the purpose of abstracting the specifics of the underlying hypervisor and improve interoperability.

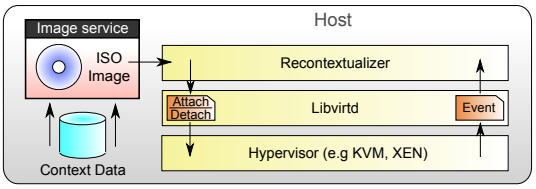

Fig. 4. Architecture overview.

4.3 Evaluation

A series of tests to evaluate the feasibility of the approach have been performed. For all tests, Libvirt version 0.9.9 is used to monitor and manage the VMs. QEMU-KVM version 1.0.50 and Xen version 4.0.0 are used as hypervisors, both running on the same hardware using CentOS 5.5 (final) with kernel version 2.6.32.24. The hosts used in these tests are on the same subnet, have shared storage and are comprised of a quad core Intel Xeon X3430 CPU @ 2.40GHz, 4GB DDR3 @ 1333MHz, 1GBit NIC and a 250GB 7200RPM WD RE3 HDD.

The results of the evaluation are shown in Figure **??**. The first set of bars illustrate the time to migrate a VM from one host to another with recontextualization running and context data attached, and the second set of columns illustrate the same migrations with recontextualization turned off and no virtual devices mounted. The third column illustrates the time spent within the recontextualizer software during the tests from the first column, measured from when the event for migration is received in the recontextualizer until the devices has been removed and reattached. The values shown are the averages from ten runs, and all columns have error bars with the (marginal) standard deviations which are all in the 0.03 to 0.07s range.

Based on the evaluation we conclude that the recontextualization process adds about an 18% overhead using either hypervisor compared to doing normal migrations. For KVM, most of the extra time required for recontextualization is spent outside the bounds of our component, likely associated with processing events and extra overhead imposed by preparing migration with virtual devices attached. In the case of Xen the device management functionality in Libvirt proved unreliable, and we therefore have to bypass the Libvirt API and rely on sub-process calls from the recontextualizer to Xen using the *xm* utility. This workaround is likely to increase the time needed for recontextualization in the Xen case.

There are four major phases associated with the recontextualization process. First, information about the VM corresponding to the event is resolved using Libvirt when the migration event is received. In the second phase, any current virtual contextualization device is identified and detached. Third, new contextualization information is prepared and bundled into a virtual device (ISO9660) image. Finally, the new virtual device is attached to the VM. A detailed breakdown of the time spent in different phases of recontextualization is presented in Figure **??**. The above mentioned workaround for Xen interactions affects the second and fourth phase (detaching and attaching of devices), most likely increasing the time required for processing. In the first and third phases Xen requires significantly longer time than KVM despite the VMs being managed using the same calls in the Libvirt API, indicating performance flaws either in the link between Libvirt and Xen or in the core of Xen itself.

4.4 Practical Experiences

When creating the system a vast number of bugs and shortcomings both with Libvirt and the underlying hypervisors were experienced. It turned out that

migrating VMs using KVM with USB devices attached periodically caused the migration to fail without any indicative error of the root cause. It was discovered after looking through the source code of qemu-kvm (due to a lack of documentation) that support for migration with USB devices is still to be fully implemented. To overcome this issue the use of a virtual CD-ROM device as a replacement was explored. Unfortunately this approach has the drawback of needing the guest OS to be configured to automatically re-mount the ISO image. We used *autofs* within our Debian guest VM for the purpose of testing. For other operating systems such as Windows that natively support the automatic mounting of CD-ROMs this would not be a problem. Using this device type in our system worked with KVM but Xen would not reliably release media mounted within a VM, causing the recontextualizer to fail in its attempt to provide new context data. To combat this issue we forced the removal of the entire CD-ROM device, reattaching another with a different ISO image.

In addition to the above, two further issues were experienced with Libvirt's support of Xen regarding events and the detaching of devices. We initially tried to use a Hardware-assisted Virtualization (HVM) guest but found that Libvirt would not propagate any VM events from the hypervisor through its API. After discovering this issue we tried using a Paravirtualised (PV) Xen guest but found that only start and stop events were available. This has had the negative effect of altering the logic of the recontextualizer, where by detaching and attaching devices incurs unnecessary overhead when a virtual machine starts, while for KVM this only occurs after migration. At the same time an issue was experienced with forcing the detachment of mounted devices in Xen via the Libvirt API. Support for the "–force" flag is available in Libvirt but the underlying implementation does not pass this to the hypervisor. To overcome this problem a workaround was implemented to issue the detach command directly to Xen via system calls to the *xm* command.

5 Conclusion and Future Work

We have shown that recontextualization is a feasible solution to the problem of using multiple cloud sites concurrently. This has been achieved by selecting an appropriate mechanism and evaluating its implementation. Our approach, based on automatic mounting of dynamically generated images as virtual devices, is highly interoperable supporting a variety of hypervisors and virtually all guest operating systems. Apart from CD-ROM mounting routines, which are standard in most operating systems, no custom software is required inside the guest VM to make the contextualization data available.

Future work includes creating a unified mechanism for contextualization and recontextualization and integrating the solution with major software projects. In addition, recontextualization mechanisms for the dynamic binding of PaaS APIs will be explored. Finally, further studies and improvements on the suggested approach will be evaluated to reduce the overhead imposed by recontextualization.

6 Acknowledgments

The research that led to these results is partially supported by the European Commission's Seventh Framework Programme (FP7/2001-2013) under grant agreement no. 257115 (OPTIMIS). We would also like to thank Tomas Forsman for technical assistance and expertize.

References

1. Armstrong, D., Djemame, K., Nair, S., Tordsson, J., and Ziegler, W. Towards a contextualization solution for cloud platform services. In *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on* (2011), IEEE, pp. 328–331.
2. Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., and Warfield, A. Xen and the art of virtualization. *SIGOPS Oper. Syst. Rev. 37*, 5 (2003), 164 – 177.
3. Bradford, R., Kotsovinos, E., Feldmann, A., and Schiöberg, H. Live wide-area migration of virtual machines including local persistent state. In *Proceedings of the 3rd international conference on Virtual execution environments* (June 2007), ACM, pp. 169–179.
4. Clark, C., Fraser, K., Hand, S., Hansen, J., Jul, E., Limpach, C., Pratt, I., and Warfield, A. Live migration of virtual machines. In *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2* (May 2005), USENIX Association, pp. 273–286.
5. Clayman, S., Galis, A., Chapman, C., Toffetti, G., Rodero Merino, L., Vaquero, L., Nagin, K., and Rochwerger, B. Monitoring Service Clouds in the Future Internet. In *Towards the Future Internet - Emerging Trends from European Research* (Amsterdam, The Netherlands, 2010), IOS Press, pp. 115–126.
6. Ferrer, A., Hernández, F., Tordsson, J., Elmroth, E., Ali-Eldin, A., Zsigri, C., Sirvent, R., Guitart, J., Badia, R., Djemame, K., Ziegler, W., Dimitrakos, T., Nair, S., Kousiouris, G., Konstanteli, K., Varvarigou, T., Hudzia, B., Kipp, A., Wesner, S., Corrales, M., Forgó, N., and Sharif, T. amd Sheridan, C. OPTIMIS: a holistic approach to cloud service provisioning. *Future Generation Computer Systems* (2011).
7. Katsaros, G., Gallizo, G., Kübert, R., Wang, T., Oriol Fito, J., and Henriksson, D. A Multi-level Architecture for Collecting and Managing Monitoring Information in Cloud Environments. In *CLOSER 2011: International Conference on Cloud Computing and Services Science (CLOSER)* (Noordwijkerhout, The Netherlands, May 2011).
8. Keahey, K., and Freeman, T. Contextualization: Providing One-Click Virtual Clusters. In *Proceedings of the 4th IEEE International Conference on eScience (ESCIENCE '08)* (Washington, DC, USA, 2008), IEEE, pp. 301 – 308.
9. Kephart, J., and Chess, D. The vision of autonomic computing. *Computer 36*, 1 (2003), 41–50.
10. Kivity, A., Kamay, Y., Laor, D., Lublin, U., and Liguori, A. kvm: the Linux virtual machine monitor. In *Proceedings of the Linux Symposium* (2007), vol. 1, pp. 225–230.
11. libvirt development team. Libvirt: The virtualization API. `http://libvirt.org/`, February 2012.

12. QEMU DEVELOPMENT TEAM. QEMU - An open source machine emulator and virtualizer. `http://www.qemu.org`, February 2012.
13. TURNBULL, J. *Pulling strings with puppet: configuration management made easy.* Springer, 2008.
14. WOOD, T., RAMAKRISHNAN, K. K., SHENOY, P., AND VAN DER MERWE, J. CloudNet: dynamic pooling of cloud resources by live WAN migration of virtual machines. In *Proceedings of the 7th ACM SIGPLAN/SIGOPS international conference on Virtual execution environments* (2011), ACM, pp. 121–132.