# Building Dynamic Computing Infrastructures over Distributed Clouds

Pierre Riteau

# Building Dynamic Computing Infrastructures over Distributed Clouds

Pierre Riteau
*University of Rennes 1, IRISA*
*INRIA Rennes - Bretagne Atlantique*
*Rennes, France*
*Email: Pierre.Riteau@irisa.fr*

***Abstract*—The emergence of cloud computing infrastructures brings new ways to build and manage computing systems, with the flexibility offered by virtualization technologies. In this context, this PhD thesis focuses on two principal objectives. First, leveraging virtualization and cloud computing infrastructures to build distributed large scale computing platforms from multiple cloud providers, allowing to run software requiring large amounts of computation power. Second, developing mechanisms to make these infrastructures more dynamic. These mechanisms, providing inter-cloud live migration, offer new ways to exploit the inherent dynamic nature of distributed clouds.**

***Keywords*-cloud computing; distributed computing; dynamic infrastructures; live virtual machine migration;**

## I. Introduction

The recently renewed interest in virtualization makes it a fundamental block for building new computing infrastructures. By decoupling execution environments from physical hardware, virtualization technologies offer more flexibility for managing computing systems. For instance, live migration of virtual machines allows to improve resource usage, increase energy efficiency, provide consolidation, etc.

This popularity of virtualization has paved the way for the advent of the cloud computing model. In cloud computing infrastructures, providers make use of virtualization technologies to offer flexible, on-demand provisioning of resources to customers. Usage of these resources is charged on a pay-for-use model. This model is also adopted for management of internal infrastructures, known as private clouds. Combining both public and private infrastructures creates so-called hybrid clouds, allowing companies and institutions to manage their computing infrastructures in flexible ways and to dynamically take advantage of externally provided resources.

Considering the growing needs for large computation power and the availability of a growing number of clouds distributed over the Internet and across the globe, this PhD thesis focuses on two principal objectives:

1) leveraging virtualization and multiple cloud computing infrastructures to build distributed large scale computing platforms,
2) developing mechanisms to make these infrastructures more dynamic – thereby offering new ways to exploit the inherent dynamic nature of distributed clouds,

This article is organized as follows. First, we present how we build large scale computing infrastructures by harnessing resources from multiple distributed clouds. Then, we describe the different mechanisms we developed to allow efficient inter-cloud live migration, which is a major building block for taking advantage of the dynamic nature of distributed clouds. Finally, we discuss ongoing and future works based on or extending these systems, and we conclude.

## II. Large Scale Computing Infrastructures over Distributed Clouds

Building large scale computing infrastructures over distributed platforms presents a number of challenges. First, the distributed aspects bring many issues such as cloud interoperability, network latency, fault tolerance, etc. Second, creating infrastructures with hundreds or thousands of nodes present new challenges linked to scalability of cloud infrastructures and distributed applications.

In this work, which started a collaboration with the Nimbus project [10] team from Argonne and University of Chicago, and researchers from the ACIS laboratory at University of Florida, we experimented with the creation of large scale virtual clusters spanning multiple distributed clouds. These clouds were built using two experimental testbeds: FutureGrid in the USA and Grid'5000 [2] in France.

To build these large scale computing infrastructures, we follow the Sky Computing [5] approach proposed to federate resources from multiple clouds. We leverage the Nimbus [10] IaaS cloud toolkit and the ViNe [17] virtual network overlay. Nimbus is used to offer a common interface across all distributed clouds, allowing the same customized execution environment to be run everywhere. We also rely on the contextualization services of Nimbus to deploy and configure these virtual clusters without manual intervention. ViNe is used as a virtual network to allow all-to-all communications between resources from multiple clouds, avoiding issues of firewalling, private IP addressing and NAT. All-to-all connectivity is a common requirement in many scientific applications.

By executing the MapReduce version of the BLAST bioinformatics application in virtual Hadoop clusters built on top of multiple distributed clouds, we showed that it is possible to efficiently run scientific applications on top of distributed cloud-based infrastructures. Of course, the level of scaling depends on the type of applications: embarrassingly parallel applications are the most suited for executing on a distributed infrastructure.

We also exploited the extension capabilities of Hadoop to dynamically adjust the virtual cluster size. This advocates that execution frameworks supporting resource addition and removal at run time are suitable to take advantage of the dynamic nature of distributed cloud computing infrastructure.

To improve the efficiency of large scale virtual cluster creation, we developed new virtual machine image deployment mechanisms for the Nimbus IaaS cloud toolkit [10]. These mechanisms leverage two kind of technologies. First, a broadcast chain mechanism (based on the Kastafior software developed at INRIA) is used to efficiently distribute virtual machine data to many physical resources. Second, a mechanism based on copy-on-write images allows near-instant virtual machine creation – radically speeding up the startup time of virtual clusters. The results of this work has been presented at the TeraGrid 2010 poster session [14].

This work did not limit itself to show that it is conceptually feasible to build large scale virtual cluster over distributed cloud computing infrastructures, but more importantly that it can be done efficiently and provide a good level of performance for the application.

## III. Leveraging the Dynamic Nature of Distributed Computing Infrastructures

When creating computing infrastructures on top of multiple distributed clouds, which may include both public and private clouds, there is a lot of opportunities for dynamic resource management. In this context, this PhD thesis proposes new mechanisms to take advantage of the dynamic nature of distributed clouds, focusing on inter-cloud live migration of virtual machines.

Live virtual machine migration [3], [9] is a powerful tool for computing infrastructure management. It allows for efficient load balancing, increases power efficiency and makes infrastructure maintenance transparent to users. However, current implementations are limited to local area networks for two reasons:

1) The migrated virtual machine cannot cross local area network boundaries without losing opened network connections,
2) Virtual machine images needs to be stored on shared file systems, usually not accessible across different data centers.

State of the art systems allow to use live migration over WANs by migrating storage and network connections [1], [20]. However, the large amounts of data to migrate make live migration over WANs expensive to use, especially when considering migrations of virtual clusters rather than single VM instances.

### *A. Shrinker: Improving Live Migration of Virtual Clusters over WANs with Distributed Data Deduplication and Content-Based Addressing*

Previous work has shown that VMs running identical or similar operating systems have a significant portion of their memory containing the same data [4], [8], [19], [21]. This is caused by VMs having the same versions of programs, shared libraries or kernels loaded in memory, or common files loaded in buffer cache. Similarly, virtual machine images contain large amounts of identical data [7], [11], [12].

The presence of this identical data has been leveraged by Sapuntzakis et al. [15] to improve virtual machine migration. However, their work predates live migration and supported only suspend/resume migration, where the VM is paused before being migrated. Additionally, they only took advantage of data available on the destination node, limiting the potential to find identical data. A similar approach was also proposed by Tolia et al. [16].

Shrinker aims to bring this mechanism in the context of live migration of virtual clusters between clouds interconnected by wide area networks. Data similarity is exploited throughout all virtual machines of the migrated virtual cluster, both in memory and on disk. Since many or all nodes composing a virtual cluster are usually based on the same operating system and run similar applications, high inter-VM data similarity can be found. Through the use of cryptographic hash functions, identical data is detected and transferred only once over the wide area link. Hash digests of duplicated pages are sent instead of the page content, which drastically reduces bandwidth utilization of the wide area link (and, consequently, migration time).

Reducing bandwidth utilization is particularly important in distributed clouds since customers are billed for incoming and outgoing network traffic. We implemented Shrinker as a modification of the KVM [6] hypervisor. Initial experiments on the Grid'5000 testbed with an implementation supporting detection of inter-VM data similarity only in memory showed that Shrinker is able to reduce migration time by 20% and wide area bandwidth usage of migration by 30 to 40% depending on workload.

A complete description of an earlier version of this work is available as a research report [13], and an improved version has been submitted to an international conference.

### *B. Network-transparency for Inter-Cloud Live Virtual Machine Migration*

A major issue in the current mechanisms for live migration is the lack of support for live migration across local area network boundaries. When a virtual machine is migrated to a

different network, all its current TCP connections are broken and it needs to be reconfigured with a new IP address to be able to use network communications again.

In the context of our collaboration with the ACIS laboratory at University of Florida, we developed mechanisms to support virtual network reconfiguration to support live migration between clouds. This work was performed within ViNe, a virtual network implementation providing all-to-all communication to nodes in grid or clouds environments, even in the presence of firewalls, NAT or private IP addressing. We modified ViNe to reconfigure itself when virtual machine mobility was detected, so that communications can remain uninterrupted. Our approach is based on standard networking techniques such as ARP proxy and gratuitous ARP messages, and leverages the ViNe infrastructure to establish tunnels between multiple cloud infrastructures. Compared to related work, this proposal focuses on the transparent detection of migrated virtual machines and how to reconfigure the virtual network overlay. More details about this work have been published at the IEEE MENS 2010 workshop [18].

### C. Autonomic Adaptation of Distributed Applications in Cloud Federations

In a federation of distributed clouds where inter-cloud live migration is available (through the work described in sections III-A and III-B), it would be possible to dynamically adapt distributed applications during run time, by relocating virtual machines to different clouds. This adaptation can be executed for several reasons:

1) Changes in resource availability: it can be interesting to migrate virtual machines to a faster cloud, or back to a private cloud when more resources were made available.
2) Changes in resource cost. Although it is not currently very developed, we envision that in the future clouds will be much more dynamic in their price (depending on current load, energy cost, etc.). As a matter of fact, Amazon already introduced some price variability in Amazon EC2 with spot instances.
3) Changes in application requirements: for instance, applications that support job deadlines may want to modify their resource requirements when deadlines are changed, which could trigger inter-cloud live migration.

This would not only enable migrating a virtual cluster from one cloud to another, but also relocating subsets of a virtual cluster. However, this kind of relocation needs to take into account communication patterns to limit communications crossing cloud boundaries. This is required for two reasons. First, network traffic between different clouds is subject to much higher latency than local networks, which could severely degrade application performance. Second, communications between different cloud infrastructures are subject to billing for the user.

This work focused on building a transparent framework using network packet capture at the hypervisor level in order to infer communication patterns in a virtual cluster. Through experiments, we showed that our framework is able to detect communication traces similar to state of the art solutions that use more invasive techniques such as library modification. This is the first step in building an autonomic adaptation system taking into account communication patterns of distributed applications.

## IV. Remaining objectives

As a more advanced application for building large scale distributed infrastructures over multiple clouds, we are working on implementing an Elastic MapReduce service harnessing resources from distributed clouds. This service will support dynamic addition and removal of virtual nodes as well as policies for resource selection. We also plan to study how job deadlines can be included in this model to perform intelligent resource selection.

By leveraging all our previous work, it is possible to migrate virtual machines between clouds. However, this is currently supported only at the hypervisor interface level. We are working on adding support for live migration at the cloud API level, in the Nimbus IaaS cloud toolkit. We are also studying the security issues that appear when live migration is used between multiple infrastructures controlled by different organizations. This mechanism will provide the necessary authentication and build a secure connection between hypervisors to allow live migration without intrusion in the destination cloud. This mechanism would allow to introduce a new kind of resources: migratable spot instances which, instead of being killed when their resource allocation is canceled, are allowed to migrate to a different cloud.

Finally, we plan to federate all these systems into a unified infrastructure framework leveraging inter-cloud live migration to autonomically adapt applications to changes in the environment.

## V. Conclusion

Virtualization technologies and cloud computing offer new ways to manage computing resources by providing more flexibility. The availability of a growing number of cloud computing platforms offers users with the possibility to answer their increasing computing power needs. In this context, this PhD thesis focuses on two principal objectives. First, we leverage virtualization and cloud computing infrastructures to build distributed large scale computing platforms, allowing to run software requiring large amounts of computation power. We validated this approach by creating large scale virtual clusters using resources from two experimental testbeds located in the USA and in France.

Second, we develop mechanisms to make these infrastructures more dynamic. We proposed Shrinker, a mechanism to decrease bandwidth usage and migration time of inter-cloud live migration of virtual clusters. We integrated reconfiguration mechanisms in ViNe, a virtual network overlay, to transparently manage inter-cloud live migration at the network level. Finally, we built a framework to detect communication patterns in distributed applications in order to guide live migration decisions. These building blocks are being integrated into higher level tools allowing users to build dynamic computing infrastructures over distributed clouds.

## Acknowledgments

This PhD thesis is carried out under the supervision of Dr. Thierry Priol and Dr. Christine Morin, senior researchers at INRIA Rennes - Bretagne Atlantique.

## References

[1] R. Bradford, E. Kotsovinos, A. Feldmann, and H. Schiöberg. Live wide-area migration of virtual machines including local persistent state. In *Proceedings of the 3rd international conference on Virtual Execution Environments (VEE '07)*, pages 169–179, 2007.

[2] F. Cappello, E. Caron, M. Dayde, F. Desprez, Y. Jegou, P. Primet, E. Jeannot, S. Lanteri, J. Leduc, N. Melab, G. Mornet, R. Namyst, B. Quetier, and O. Richard. Grid'5000: A large scale and highly reconfigurable grid experimental testbed. *IEEE/ACM International Workshop on Grid Computing*, pages 99–106, 2005.

[3] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live Migration of Virtual Machines. In *Proceedings of the 2nd Symposium on Networked Systems Design & Implementation (NSDI '05)*, pages 273–286, 2005.

[4] D. Gupta, S. Lee, M. Vrable, S. Savage, A. C. Snoeren, G. Varghese, G. M. Voelker, and A. Vahdat. Difference Engine: Harnessing Memory Redundancy in Virtual Machines. In *8th USENIX Symposium on Operating Systems Design and Implementation (OSDI '08)*, pages 309–322, 2008.

[5] K. Keahey, M. Tsugawa, A. Matsunaga, and J. Fortes. Sky computing. *IEEE Internet Computing*, 13:43–51, September 2009.

[6] A. Kivity, Y. Kamay, D. Laor, U. Lublin, and A. Liguori. kvm: the Linux Virtual Machine Monitor. In *Proceedings of the 2007 Linux Symposium*, volume 1, pages 225–230, June 2007.

[7] A. Liguori and E. V. Hensbergen. Experiences with Content Addressable Storage and Virtual Disks. In *Proceedings of the First Workshop on I/O Virtualization (WIOV '08)*, 2008.

[8] G. Milos, D. G. Murray, S. Hand, and M. Fetterman. Satori: Enlightened Page Sharing. In *Proceedings of the 2009 USENIX Annual Technical Conference (USENIX '09)*, 2009.

[9] M. Nelson, B.-H. Lim, and G. Hutchins. Fast Transparent Migration for Virtual Machines. In *Proceedings of the 2005 USENIX Annual Technical Conference (USENIX '05)*, pages 391–394, 2005.

[10] Nimbus Project. Nimbus. http://www.nimbusproject.org/.

[11] N. Partho, M. A. Kozuch, D. R. O'Hallaron, J. Harkes, M. Satyanarayanan, N. Tolia, and M. Toups. Design tradeoffs in applying content addressable storage to enterprise-scale systems based on virtual machines. In *Proceedings of the 2006 USENIX Annual Technical Conference (USENIX '06)*, pages 1–6, 2006.

[12] S. Rhea, R. Cox, and A. Pesterev. Fast, inexpensive content-addressed storage in foundation. In *Proceedings of the 2008 USENIX Annual Technical Conference (USENIX '08)*, pages 143–156, 2008.

[13] P. Riteau, C. Morin, and T. Priol. Shrinker: Efficient Wide-Area Live Virtual Machine Migration using Distributed Content-Based Addressing. Research Report RR-7198, INRIA, February 2010.

[14] P. Riteau, M. Tsugawa, A. Matsunaga, J. Fortes, T. Freeman, D. Labissoniere, and K. Keahey. Sky Computing on FutureGrid and Grid'5000. In *TeraGrid'10 (Poster session)*, 2010.

[15] C. P. Sapuntzakis, R. Chandra, B. Pfaff, J. Chow, M. S. Lam, and M. Rosenblum. Optimizing the migration of virtual computers. In *Proceedings of the 5th symposium on Operating systems design and implementation (OSDI '02)*, pages 377–390, 2002.

[16] N. Tolia, T. Bressoud, M. Kozuch, and M. Satyanarayanan. Using Content Addressing to Transfer Virtual Machine State. Technical report, Intel Corporation, 2002.

[17] M. Tsugawa and J. Fortes. A virtual network (ViNe) architecture for grid computing. In *International Parallel and Distributed Processing Symposium*, 2006.

[18] M. Tsugawa, P. Riteau, A. Matsunaga, and J. Fortes. User-level Virtual Networking Mechanisms to Support Virtual Machine Migration Over Multiple Clouds. In *The 2nd IEEE International Workshop on Management of Emerging Networks and Services (IEEE MENS 2010)*, 12 2010.

[19] C. A. Waldspurger. Memory resource management in VMware ESX server. In *Proceedings of the 5th symposium on Operating systems design and implementation (OSDI '02)*, pages 181–194, 2002.

[20] T. Wood, K. Ramakrishnan, P. Shenoy, and J. van der Merwe. CloudNet: Dynamic Pooling of Cloud Resources by Live WAN Migration of Virtual Machines. In *Proceedings of the 2011 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE '11)*, 2011.

[21] T. Wood, G. Tarasuk-Levin, P. Shenoy, P. Desnoyers, E. Cecchet, and M. Corner. Memory Buddies: Exploiting Page Sharing for Smart Colocation in Virtualized Data Centers. In *Proceedings of the 2009 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE '09)*, pages 31–40, 2009.