Performance Variation of the Public Cloud

Guilherme G. Haetinger, Eduardo Roloff, Philippe O. A. Navaux
Informatics Institute
Federal University of Rio Grande do Sul - UFRGS
Porto Alegre, Brazil
{gghaetinger,eroloff,navaux}@inf.ufrgs.br

Abstract

Cloud Computing became a mainstream option as an environment to execute applications of any type. In recent years a lot of effort was made by the community to migrate High-Performance Computing (HPC) applications to the Cloud. However, the cloud still presents some drawbacks that need to be addressed. Aspects like network interconnection, raw performance, and high variability are the major barriers to the adoption of the cloud in large scale by the HPC community. In this work, we investigate the performance variation of Virtual Machines (VM) among different datacenters of the Microsoft Azure cloud provider. In our experiments, we observed that the same VM type presented a performance variation of 1.46%. Large performance variations are not desirable for HPC purposes, and this remains as an interesting research topic of cloud computing.

1. Introduction

Cloud Computing public providers normally offer a large number of locations across the world. Research in High-Performance Computing (HPC) using the cloud normally focuses on performance evaluation [6, 7, 8] and application migration [1, 10, 2]. However, there is no guarantee from the provider side that an instance of the same type and size will present the same performance over all datacenter locations. Due to this, one important aspect that needs investigation is the performance variation of the same instance among different datacenters [9, 5].

In this work, we present a performance evaluation of the Microsoft Azure [1] cloud provider, using four different locations. We observed that, even using exactly the same instance type, there is a performance variation of up to 1.46%(at maximum measurement level) when comparing the best and worst datacenter.

1 https://azure.microsoft.com

The rest of this paper is organized as follow. The Section 2 provides a brief summary of the Microsoft Azure datacenters and motivates our work. In Section 3 we describe what our hypothesis is and how we verified it. Section 4 shows and explains our results so far. Finally, in Section 5 we discuss the future work and the next steps that will be made in our research.

2. Background

Microsoft Azure is a cloud platform that has datacenters all over the globe so that people can use them for storage and machine virtualization. They charge for what you use, in a monthly subscription, where price is relative to the machine's characteristics, having a price range from $13 - $2,000.

2.1. Instance Types

There are different machine specifications in their inventory. The main ones would be:

A0-4 (Basic): 'A' Basic is an economically simple option for cloud development and storage.

A8-11 (High Performance): A8-11 instances feature Intel® Xeon® E5 processors, used for parallel programming and, consequently, perfect for cluster workloads development and MPI execution.

Av2 (Standard): Av2 Standard is the latest generation of A series virtual machines with similar CPU performance and faster disk, being more suitable for servers and web services.

D2-64 (New Generation): D2-64 v3 instances are the latest generation of General Purpose Instances.

Azure has more machines types, a summary of instance variations of Azure can be found in Table 1.

Designation	#Inst.	Price range
General Purpose	27	0.018 – 3.200
Compute Optimized	5	0.060 – 0.997
Memory Optimized	20	0.148 – 8.690
Storage Optimized	4	0.344 – 2.752
GPU/FPGA	7	0.900 – 4.370
HPC	6	0.971 – 2.136

Table 1. Instances characteristics of Microsoft Azure in West USA datacenters (verified on July 25, 2017). Price ranges are given in US$/h.

2.2. Azure locations

Microsoft Azure has 26 datacenters active(2017) in 10 different countries that are available for regular users, as well as some datacenters that are suitable only for private usage (like the Department of Energy exclusive datacenter).

The locations where Azure is available are: Australia (2 sites), Brazil (1 site), Canada (2 sites), India (3 sites), USA (8 sites), Asia-Singapore(2 sites), Japan (2 sites), South Korea (2 sites), Europe (2 sites) and UK (2 sites).

3. Motivation and Proposal

Because of the large number of datacenters locations detailed in Section 2, we decided to research whether there are any differences between these datacenters' performances, since low performance variation is an important aspect for HPC [4]. In case we find any dissimilarity, we intend to discover why and how we can use that.

4. Evaluation

This section describes the methodology used in our experiments. The results are presented and described as well.

4.1. Methodology

To evaluate the datacenters' machines efficiency, we needed a benchmark to measure the processing power of each of the VM computer's core. For that reason we created a benchmark, using C programming language and OpenMP [3], to execute a test in parallel in every processing core. This benchmark, named Leveled Core Individual Efficiency Benchmark (LCIE), consists in generating random integers according to the load level selected for each core. The purpose of the LCIE benchmark is to simulate simply real-life applications that present a load imbalance and not constant mathematical algorithms. It was designed to simulate different loads using the available cores of the environment. The benchmark is still a work in progress, but the first prototype was used in this work.

To get the needed results for the research, we executed a benchmark with the given proportion of '100000' five times in a single VM in each of the following USA regions: West, East, South and North. We executed it in a A8 virtual machine with Linux Ubuntu 64-bits as the OS and these specifications:

- 8 cores;
- 56GB of RAM;
- 382GB of storage;

Each A8 machine has 8 cores, and we used our benchmark to simulate four load levels of an application. The loads that we simulated represent proportions of 100%, 75%, 50% and 25% in terms of application instructions. Each load level was executed by 2 cores in every execution so we could fit all cores in the benchmark.

4.2. Results

The Figure 1 shows the results of our experiments. We executed our benchmark with four different load levels, to simulate an application with load imbalance among its processes. Considering what was explained in subsection 4.1, the four levels simulate workloads of 25%, 50%, 75% and 100% and are represented in the x axis. The execution time is represented in the Y axis. The error level is plotted in the figure as well, and the variation was low, which means that our results are valid.

It is possible to see in the figure that we have performance differences in all levels. The best overall location was the East, performing better than all the other three location in all the load levels. The second best location, that is close to the East, was the South. The West and North locations presented less performance than East and South. One important thing that we want to remark is the 75% load level for the West and North locations. West was slightly faster than North in all other levels, but in the 75% level, the North surpasses the West. This means that we observed performance differences according with the usage of the cores.

The Table 2 shows the performance gains or losses of all the location for all the levels compared to the West location. As we can see, there is a variation among the levels and their order of efficiency is not constant at all, specially the 100% level, in which the improvements of the East and South locations in relation with the others presented a huge decrease of the performance gain and we could not determine a pattern. The North location presented a chaotic behavior, taking the worst efficiency place in most of the times.

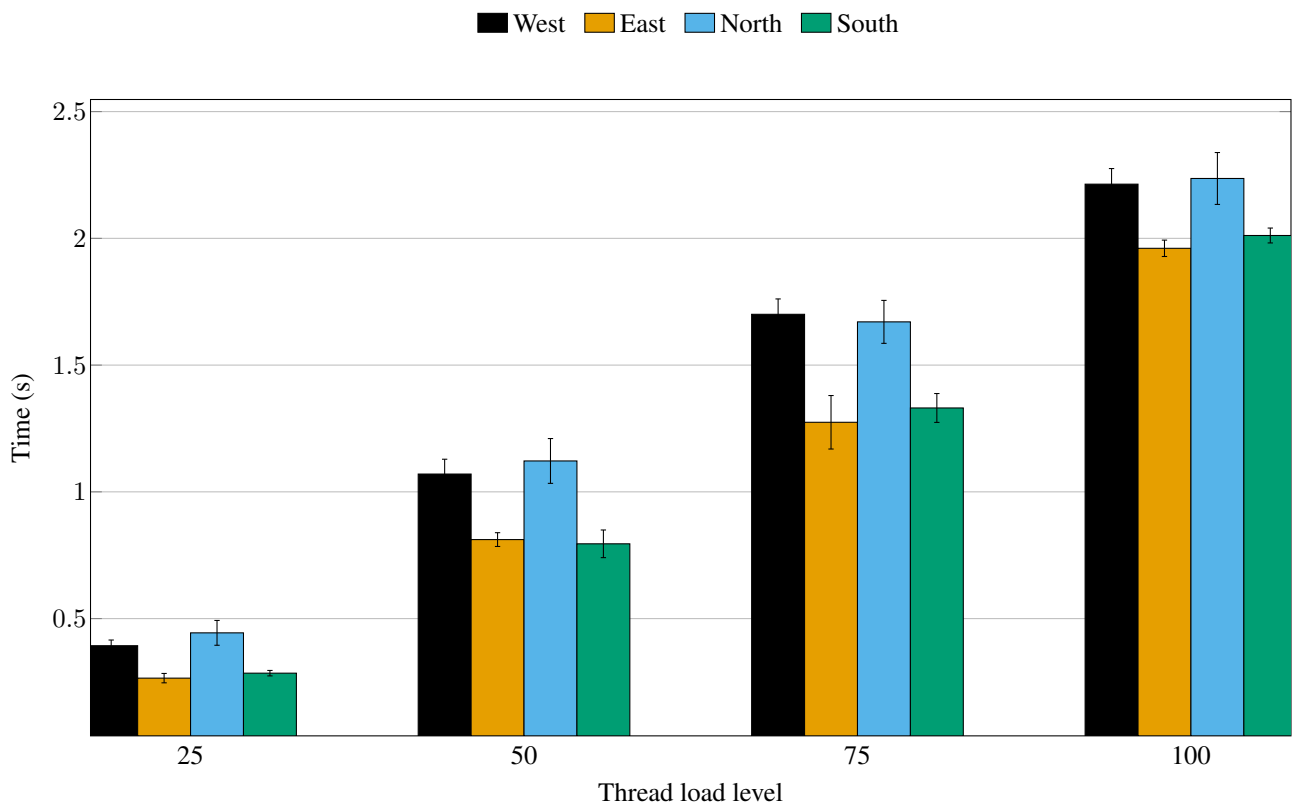

Figure 1. Performance Results for the USA main regions

As we gather the results in Figure 1, we can conclude that there is a significant difference between the instances from different locations. This performance variance is not desirable for HPC users, because it is important to have an stable environment.

The variance of data can affect the use from millions of public cloud users negatively by not providing the best experience possible on cloud workloads. However, both the differences of performance among these four locations and the differences observed in performance gains and losses for the different load levels are promising research topics that need further investigation.

Load Level	East	North	South
25	48.25%	-11.43%	38.03%
50	31.85%	-4.60%	34.62%
75	33.43%	1.79%	27.76%
100	12.91%	-1.00%	10.07%

Table 2. Performance gains and losses using the West location as baseline.

5. Discussion

The Cloud Computing paradigm has a tremendous potential to be the *de-facto* standard environment for all kinds of applications. Moreover, it presents unique characteristics that are very promising for HPC users, such as the capability to build a completely customized environment in few minutes.

In this article, we focus on the performance differences among the same machine configuration used in different environments. Our results proved that there is a performance difference, even when we expected the same performance, because the configuration of the instances was the same.

We introduce our performance benchmark project as well. It will be used to test diverse cloud environments by simulating applications with processes that have different loads.

As future work, we will continue to investigate the cloud as an environment to execute HPC applications, with the main focus on the view from the user's side. We will create a mechanism for profiling the environments and help the user to choose the best suitable to their needs.

In the benchmark side, we will continue to improve it to create a complete toolkit to test the cloud. The purpose is to offer a tool that helps the user to measure the efficiency

of different cloud environments to help him or her to save money in their executions. As far as we know there is not a benchmark that was designed specifically for the cloud with focus in the HPC community.

Acknowledgments. This research received funding from the EU H2020 Programme and from MCTI/RNP-Brazil under the HPC4E project, grant agreement no. 689772. Additional funding was provided by FAPERGS in the context of the GreenCloud Project.

References

[1] P. V. Beserra, A. Camara, R. Ximenes, A. B. Albuquerque, and N. C. Mendonça. Cloudstep: A step-by-step decision process to support legacy application migration to the cloud. In *Maintenance and Evolution of Service-Oriented and Cloud-Based Systems (MESOCA), 2012 IEEE 6th International Workshop on the*, pages 7–16. IEEE, 2012.

[2] E. D. Carreño, E. Roloff, and P. O. A. Navaux. Towards weather forecasting in the cloud. In *24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2016, Heraklion, Crete, Greece, February 17-19, 2016*, pages 659–663, 2016.

[3] L. Dagum and R. Menon. Openmp: an industry standard api for shared-memory programming. *Computational Science & Engineering, IEEE*, 5(1):46–55, 1998.

[4] Y. El-Khamra, H. Kim, S. Jha, and M. Parashar. Exploring the performance fluctuations of hpc workloads on clouds. In *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, pages 383–387. IEEE, 2010.

[5] S. K. Garg, S. K. Gopalaiyengar, and R. Buyya. Sla-based resource provisioning for heterogeneous workloads in a virtualized cloud datacenter. In *International conference on Algorithms and architectures for parallel processing*, pages 371–384. Springer, 2011.

[6] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema. Performance analysis of cloud computing services for many-tasks scientific computing. *IEEE Transactions on Parallel and Distributed systems*, 22(6):931–945, 2011.

[7] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema. A performance analysis of ec2 cloud computing services for scientific computing. In *International Conference on Cloud Computing*, pages 115–131. Springer, 2009.

[8] E. Roloff, M. Diener, A. Carissimi, and P. O. Navaux. High performance computing in the cloud: Deployment, performance and cost efficiency. In *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*, pages 371–378. IEEE, 2012.

[9] A. Shieh, S. Kandula, A. G. Greenberg, and C. Kim. Seawall: Performance isolation for cloud datacenter networks. In *HotCloud*, 2010.

[10] V. Tran, J. Keung, A. Liu, and A. Fekete. Application migration to cloud: a taxonomy of critical factors. In *Proceedings of the 2nd international workshop on software engineering for cloud computing*, pages 22–28. ACM, 2011.