Understanding the Cost of the Cloud for Scientific Applications

Iman Sadooghi
Computer Science
Illinois Institute of Technology
Chicago, IL, USA
isadoogh@iit.edu

Ioan Raicu
Computer Science
Illinois Institute of Technology
Chicago, IL, USA
iraicu@cs.iit.edu

ABSTRACT

Commercial clouds bring a great opportunity to the scientific computing area. Scientific applications usually need huge resources to run on, however not all of the scientists have access to significant high-end computing systems, such as those found in the Top 500 list. Cloud has gained the attention of scientists as a competitive resource to run HPC applications at a lower cost. But as a different infrastructure, it is unclear whether clouds are capable of running scientific applications with a reasonable performance. Before we can start using existing public cloud platforms for scientific or in general, high I/O demanding applications, we have to study the raw performance of public clouds in terms of compute, memory, network and I/O.

We assess the ability and the cost of the Amazon EC2 cloud running scientific applications using customized instances against the local systems with no virtualization. The paper develops a full set of metrics and conducts a comprehensive evaluation over Amazon EC2 in the following aspects: we measure the performance of memory, CPU, network, and I/O for each instance type of Amazon EC2. We also create a virtual cluster to evaluate the compute and I/O performance of multiple instances and different services on Amazon. Finally, we analyze the cost of using cloud for scientific computing and try to find the most cost-effective instances in different use case scenario.

INTRODUCTION

The idea of using clouds for scientific applications has been around for several years, but it has not gained traction primarily due to many issues such as lower network bandwidth or poor and unstable performance. Scientific applications often require high performance systems that have high I/O and network bandwidth. Using commercial clouds gives scientists opportunity to use the larger resources on-demand. However, there is an uncertainty about the capability and performance of clouds to run scientific applications because of their different nature. Most of the cloud resources use commodity network with significantly lower and less stable bandwidth than supercomputers. The virtualization overhead is also another issue that leads to variable compute and memory performance. I/O is yet another factor that has been one of the main issues on application performance.

The main goal of this research is to evaluate the cost and the performance of the Amazon cloud as the most popular commercial cloud available. We run micro benchmarks and real applications on Amazon EC2 and S3 to evaluate its performance on critical metrics [1]. We also identify the weaknesses and advantages of the cloud environment on scientific computing area and present the ways to optimize the cloud performance over scientific applications.

METHODOLOGY

Our methodology to evaluate the capability of different instance types of Amazon EC2 is divided into three parts: 1. Run the micro benchmarks to measure the actual performance and compare with the theoretical peak that we expect to get. Also include a non-virtualized system, to understand virtualization effect. 2. Evaluate the performance of a virtual cluster of multiple instances, running real applications such as HPL[2]. 3. Analyze the cost of the cloud based on the actual performance results.

PERFORMANCE EVALUATION OF EC2 AND S3

Fig. 1 shows the system memory read bandwidth in different memory hierarchy levels. The vertical axis shows the cache size. The bandwidth is very stable up to a certain cache size. The bandwidth starts to drop after a certain size due to surpassing the memory cache size at a certain hierarchy level.

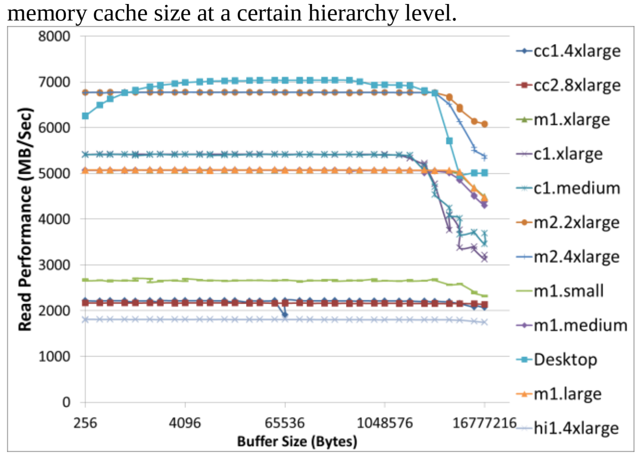

Fig. 1. CacheBench Read benchmark results, one benchmark process per instance

Fig. 2 shows the compute performance and efficiency of each instance on HPL as well as the claimed performance by Amazon.

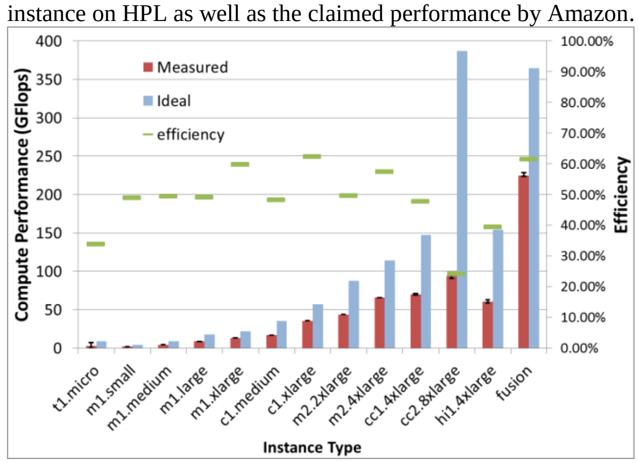

Fig. 2. HPL benchmark results: compute performance and efficiency of single instances comparing with their ideal performance.

Among the Amazon instances, the cc2.8xlarge and t1.micro have the highest and the lowest compute performance. Although the cc2.8xlarge instance has the largest compute capacity among the instances, it is the most inefficient instance. The reason for that lies behind the number of the cores in this instance. cc2.8xlarge has 16 cores. The expected performance is the aggregate performance of all of the cores of the instance. But the real performance is lower because of the communication overhead of 16 cores which is caused by the MPI application. Other papers have also reported the poor MPI performance on EC2 cloud [3][4].

Fig. 3 shows the network latency distribution of EC2 m1.small instances. It also plots the hop distance of two instances. The network latency in this experiment varies between 0.006 ms and 394 ms. There is no clear correlation between the latency and the hop distance. The high latency variance is not desirable for scientific applications.

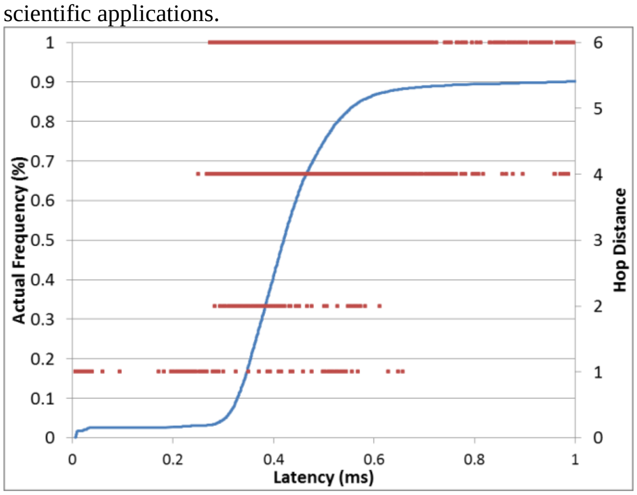

Fig. 3. Cumulative Distribution Function and Hop distance of connection latency between instances inside a datacenter.

Fig. 4 shows that the performance of the cluster of the m1.medium instances is much lower than the ideal performance, especially at large scales. The main reason is the communication overhead of instances and the MPI overhead of the HPL benchmark. The efficiency drops as the size of cluster gets bigger.

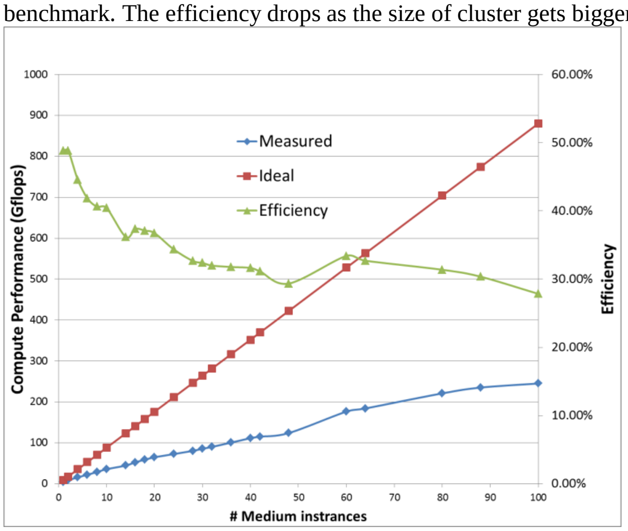

Fig. 4. Compute performance and efficiency of multiple m1.medium instances.

Fig. 5. Read throughput, of S3 compared to other systems shows the read throughput of s3 in different scales compared with two distributed file systems running on EC2. The throughput of S3 is lower. But it scales linearly. Therefore the aggregate throughput is comparable to other systems at larger scales.

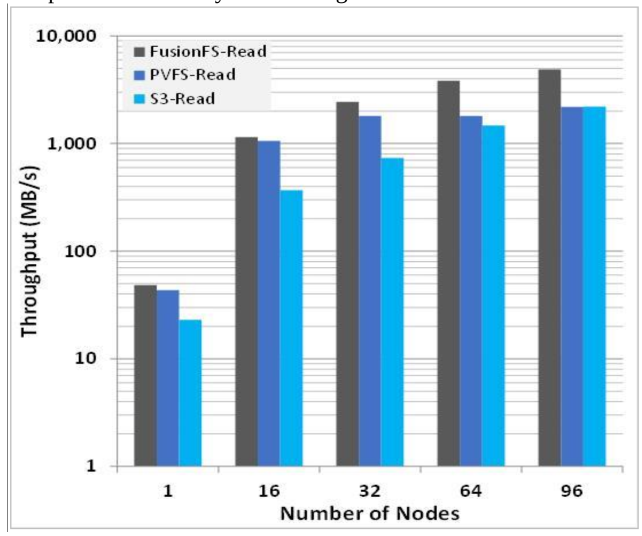

Fig. 5. Read throughput, of S3 compared to other systems

CONCLUSION AND FUTURE WORK

In this paper, we evaluate the performance of various instance types by running micro benchmarks on memory, compute, network and storage. In most of the cases, the actual performance of the instances is lower than the expected performance or what is claimed by Amazon. Most of the instances have stable memory bandwidth which is comparable with non-virtualized systems. However, we observe strange low memory performance on the HPC instances. The compute performance of the instances is affected by virtualization overhead on the larger instances. the multimode compute tests show low performance and poor efficiency of the virtual cluster on larger scales. The network latency on EC2 is higher and less stable than HPC environments.

In conclusion, we believe the Amazon EC2 instances are not yet comparable with high end HPC systems.

In our future work we would like to enhance our work on multiple cloud infrastructures including Microsoft Azure. We also want to set up our own private cloud in which we can try different open source Middleware including Eucalyptus and OpenNebula.

Another direction of this research is toward optimizing the virtualization overhead. We would like to deploy Palacios hypervisor. Palacios offers many features like Pass-through I/O technique and paging mechanisms that adds a very low virtualization overhead [5].

REFERENCES

[1] Amazon Elastic Compute Cloud (Amazon EC2), Amazon Web Services, [online] 2012, http://aws.amazon.com/ec2/

[2] A. Petitet, R. C. Whaley, J. Dongarra, A. Cleary. "HPL", [online] 2008, http://www.netlib.org/benchmark/hpl/

[3] Q. He, et al. "Case study for running HPC applications in public clouds," In *Proc. of HPDC*, 2010.

[4] Guohui Wang and T. S. Eugene Ng. The Impact of Virtualization on Network Performance of Amazon EC2 Data Center. In IEEE INFOCOM, 2010.

[5] J. Lange, K. Pedretti, P. Dinda, P. Bridges, C. Bae, P. Soltero, A. Merritt, "Minimal Overhead Virtualization of a Large Scale Supercomputer," Proceedings of the 2011 ACM