Egyptian Informatics Journal (2016) 17, 175–182

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Scalability and communication performance of HPC on Azure Cloud

Hanan A. Hassan, Shimaa A. Mohamed *, Walaa M. Sheta

Informatics Research Institute, The City for Scientific Research and Technology Applications, Egypt

Received 19 August 2015; accepted 1 November 2015
Available online 25 November 2015

KEYWORDS

HPC on cloud computing;
Azure Cloud;
NPB benchmarks

Abstract Different domains of research are moving to cloud computing whether to carry out compute intensive experiments or to store large datasets. Cloud computing services allow the users to focus on the application, without being concerned about the physical resources. HPC system on the cloud is desired for their high needs of efficient CPU computations. Our objective was to evaluate the scalability and performance of High Performance Cloud Computing on Microsoft Azure Cloud infrastructure by using well known Benchmarks, namely, IMB point-to-point communication and NAS Parallel Benchmarks (NPB). In our experiments, performance of the HPC applications on the cloud is assessed in terms of MOPS and speedup, and is tested under different configurations of cluster sizes. Also, point-to point communication performance between nodes is assessed in terms of latency and bandwidth as a function of message size.

1. Introduction

Cloud computing is typically based on sharing a set of accessible commodity computing resources, located all over the world and available on demand over a network. Various services are delivered to end users to access these resources such as: "Infrastructure as a service" (IaaS), Platform as a service (PaaS), and Software as a service (SaaS). Contrary to conventional HPC system which was mostly built on dedicated resources data centre with huge computing capacity with the difficulty to be multiplexed and expanded. Thus, many users applications could not be accommodated because the HPC facilities may not grow as fast as the rising of computational demands.

HPC applications are increasingly being used in many fields such as scientific research, business and data analytics [1].

Virtualization technology introduces attractive techniques to manage and multiplex computing resources. It contributed to the presence of cloud computing. Cloud computing has become an alternative platform to fill the gap between scientists growing computational demands and limited local computing capabilities [2]. In addition to virtualization, cloud computing has many benefits which are introduced to HPC applications users as elasticity of resources and elimination of cluster setup cost and time [1]. In virtualized HPC system, computing nodes are deployed via individual virtual machines

* Corresponding author.

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.11.001

connected over a network. Thus, capacity and structure of HPC on cloud computing are desired to be adjusted dynamically according to the requirement of customers [2].

Ethernet is the most technology used to communicate between virtual machines in current cloud computing. HPC applications require low latency and high bandwidth interprocess communication. However, existence of traditional interconnection and network virtualization on the cloud caused a performance bottleneck of HPC applications [1]. New virtualization solutions have been proposed that use kernel-based virtual machine (KVM) and XEN hypervisors to solve the performance bottleneck by reducing the virtualization management overhead and by allowing direct access from the VMs to the network [3]. In addition, several technologies have been proposed to solve communication performance between nodes. InfiniBand as an advanced interconnect technology would be a better choice because of great performance. More than 44% of the fastest supercomputing systems rely on InfiniBand for their I/O and networking requirements [1,4–6].

Many cloud providers deliver environments for developing and deploying applications in the cloud such as Amazon, Rackspace, and Microsoft Azure [7]. Microsoft Azure provides powerful compute and storage resources on demand through hardware level virtualization. It provides the possibility of computing on virtual parallel clusters. Some studies investigated the benefits of performing HPC applications on the Microsoft Azure Cloud. In spite of the benefits offered by cloud computing, it has not yet been established whether Cloud can offer a suitable alternative to supercomputers for HPC applications. Therefore, this motivated us to carry out a detailed study of HPC on the Microsoft Azure Cloud.

This paper evaluates High Performance Cloud Computing on Microsoft Azure Cloud infrastructure, the largest public cloud in production, using up to 128 total cores per cluster of VMs. In this evaluation, the scalability of representative parallel HPC codes of the NAS Parallel Benchmarks (NPB) suite [8] was studied.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 introduces the experimental configuration and methodology of the work. Section 4 analyzes the performance results of the selected message-passing middleware on cloud of Microsoft Azure. These results have been obtained from a micro-benchmarking of point-to-point primitives, as well as an application benchmarking using representative HPC codes in order to analyze the scalability of HPC applications. Section 5 gives some concluding remarks and Future Work.

2. Related work

This section is oriented toward three main subjects: (1) benchmarks and technologies that used to measure the performance of MPI message passing communication on parallel computing, (2) performance analysis of running HPC on the public cloud providers, and (3) the feasibility and the computing performance of running HPC applications on cloud platforms.

Many benchmark programs were used to measure the performance of MPI on parallel computing as SKaMPI [9], Mpptest [10], IMB [11], MPBench [12] and MPIBench [13]. Ismail et al. discuss the results of MPI message passing communication on Razi (Gigabit Ethernet) and Haitham (InfiniBand) clusters by using SKaMPI, IMB and MPBench applications. Then, they compared output results from these applications and analyzed for validation. They reached that the architecture of the clusters itself might also affect the results independent of type of interconnect [4].

Mauch et al. [6] present an architecture for HPC IaaS clouds using high speed cluster interconnects (InfiniBand) that allows an individual network configuration with QoS mechanisms. They considered an HPC cloud model that used InfiniBand in a virtualized environment instead of Ethernet devices, so as to provide lower latency in network communication for scientific applications by using the High Performance Linpack (HPL) [14] benchmark. They compared their virtual InfiniBand network to a local one, and showed that network latency is constant, of the order of 0.1−0.2 μsec, independently of the message size.

Expósito et al. [15] figured out that although the Amazon EC2 cluster compute instances provide a high-speed network (10 Gigabit Ethernet), it is still penalized by the absence of an suitable I/O virtualization support, thus preventing a full virtualization of the network interface controller (NIC). In [16], Jose et al. offer an in-depth analysis on SR-IOV with InfiniBand [17]. They evaluated the performance of MPI and PGAS point-to-point communication benchmarks over SR-IOV with InfiniBand and compared to that of the native InfiniBand hardware; for most message lengths. Their results reveal that the performance of MPI collective operations over SR-IOV with InfiniBand is inferior to native (non-virtualized) mode. Therefore, they evaluated the trade-offs of various VM to CPU mapping policies on modern multi-core architectures.

According to second category, several cloud providers interested in building HPC systems on the cloud, and measured its performance using different benchmarks. Roloff et al. [18] compared HPC applications running on three cloud providers, Amazon Ec2, Microsoft Azure and Rackspace by using NAS Parallel Benchmark [8]. In addition, they analyze three important characteristics of HPC such as deployment facilities, performance and cost efficiency. Finally, they compared them to a cluster of machines. Their results figure out that HPC in the cloud can have a higher performance and cost efficiency than a traditional cluster, up to 27% and 41%, respectively. Expósito et al. [15] assessed the performance and scalability of Virtual HPC on Amazon EC2 Cluster Compute (CC) by using NAS parallel benchmark up to 512 core. Akioka et al. used two well-known benchmarks which are NAS parallel benchmark, and high-performance linpack benchmark for distributed-memory computers (HPL) on Amazon EC2. For each requested number of cores (VCPUs), the results figure out that the execution performance greatly fluctuates among the different runs of the same configuration, especially, when the benchmark was run by a large number of requested cores [19]. Hassani et al. proposed a new approach to improve the performance and scalability of HPC application on Amazons HPC Cloud by implementing the MPI version of parallel Radix sort, analyzed its performance on cloud infrastructure and finally compared it with dedicated HPC cluster. Their results reveal a significant improvement in speedup and scale up for up to 8 nodes with the response rate of more than 20% parallel efficiency on the Cloud in comparison with dedicated HPC cluster [20].

Table 1 The relevant related work.

Authors name	Benchmarks	Performance metrics	Organization
R. Ismail, N.A.W.A. Hamid, M. Othman, R. Latip, and M.A. Sanwani [4]	SKaMPI, IMB, MPBench based on OpenMPI library	Latency and bandwidth	Razi (Gigabit Ethernet), Haitham (InfiniBand) clusters
V. Mauch, M. Kunze, and M. Hillenbrand [6]	HPL	Latency	Amazon EC using InfiniBand network
R.R. ExpóSito, G.L. Taboada, S. Ramos, J. TouriñO, and R. Doallo [15]	IMB point-to-point communication over 10 Gigabit Ethernet, and NAS parallel benchmarks	Latency and bandwidth MOPS and Speedup	Amazon EC2
J. Jose, M. Li, X. Lu, K.C. Kandalla, M.D. Arnold, and D. K. Panda [16]	MPI, PGAS point-to-point communication over SR-IOV with InfiniBand	Latency and bandwidth	–
E. Roloff, M. Diener, A. Carissimi, and P.O.A. Navaux [18]	NAS parallel benchmarks	Execution time of benchmark, cost and cost efficiency	Amazon EC2 Microsoft Azure Rackspce
S. Akioka and Y. Muraoka [19]	NAS parallel benchmark and HPL	Gflops and MOPS	Amazon EC2
R. Hassani, M. Aiatullah, and P. Luksch [20]	MPI version of parallel Radix sort	Speedup	Amazon EC2
M.B. Belgacem and B. Chopard [21]	Concurrent multiscale jobs based on MPI library	Execution time	Amazon cloud resources (USA) and MAPPER computing infrastructure (Switzerland)

Belgacem et al. connected EC2 based cloud clusters located in USA to university clusters located in Switzerland, and ran a tightly coupled, concurrent multi-scale MPI based application on this infrastructure. Then, they measured the overhead induced by extending their HPC clusters with EC2 resources. Their results show that Applying multi-scale computation on cloud resources can lead to low performance without a proper adjustment of CPUs power and workload. Nevertheless, by enforcing a load-balancing strategy one can benefit from the extra Cloud resources [21].

Several researches have been carried out to evaluate the ability and the computing performance of running HPC applications on cloud platforms. Gupta and Milojicic [1] propose that Cloud can be applicable platform for some HPC applications depending upon application characteristics such as communication volume and pattern and sensitivity to OS noise and scale. They had evaluated the performance of HPC applications on a range of platforms varying from cloud (with and without virtualization) to HPC-optimized clusters. They found that their Cloud is feasible platform for low communication intensive applications such as embarrassingly parallel and tree-structured computations and HPC-optimized clusters are better for the rest [1]. In [22], Evangelinos et al. confirmed the feasibility of running MPI based on atmosphere applications on the Amazon EC2 cloud cluster which compared its performance to low-cost cluster system. In addition, Metrotra et al. compared the execution of NASA HPC applications in both Amazon EC2 and NASA primary supercomputer "Pleiades" using a range of cores between 4 and 256. They proved that the cc1.xlarge EC2 instance type cannot quite compete with "Pleiades" [23]. Carreño et al. present the challenges of performing a numerical weather prediction (NWP) application on the Microsoft Azure Cloud platform, and compared the execution of this High-Performance Computing (HPC) application in a local cluster and the cloud using different instances sizes. Results show that cloud infrastructure can be used as an applicable HPC alternative for this application.

This variety of research efforts to build and assess HPC system on cloud computing using different types of Benchmarks and technologies can serve. Table 1 summarizes the main efforts of the most relevant publications to our proposed. These efforts are categorized according to type of benchmarks, performance metrics and organization (cloud provider). For example, authors in [4,6,15,16] proposed high speed cluster interconnects InfiniBand to improve communication performance between cluster nodes, while authors in [15,18–21] analyzed characteristics of HPC system on different cloud providers to be able to run several HPC applications on cloud platforms. In our experiments, we assessed HPC system in Microsoft Azure Cloud using the NAS Parallel Benchmarks (NPB).

3. Experimental configuration and evaluation methodology

The performance evaluation has been conducted on a varying size cluster up to 16 VMs. The instances configuration are shown in Table 2.

Widely used HPC messaging middleware such as OpenMPI [25] 1.4.4 and MPICH2 [26] 1.4.1 was selected as a running environment of native codes (C/C++ and Fortran) of NBP benchmark.

The evaluation consists of a micro-benchmarking of point-to-point process data transfer, both inter-VM (an Ethernet communication model) and intra-VM (a shared memory communication model), at the message passing library level.

Table 2 Description of the azure cluster compute [24].

Instance size	Cores	CPU type	RAM	RAM type	Operating system
A10	8	Intel Xeon E5-2670 @ 2.6 GHz	56 GB	DDR3-1600 MHz	Centos 6.5

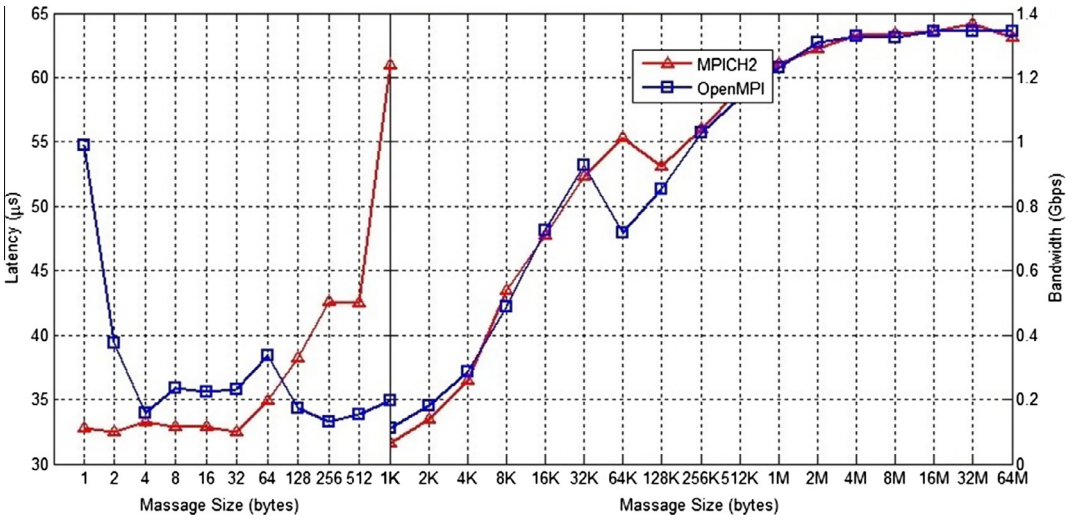

Figure 1 Point-to-point communication performance on Azure instances.

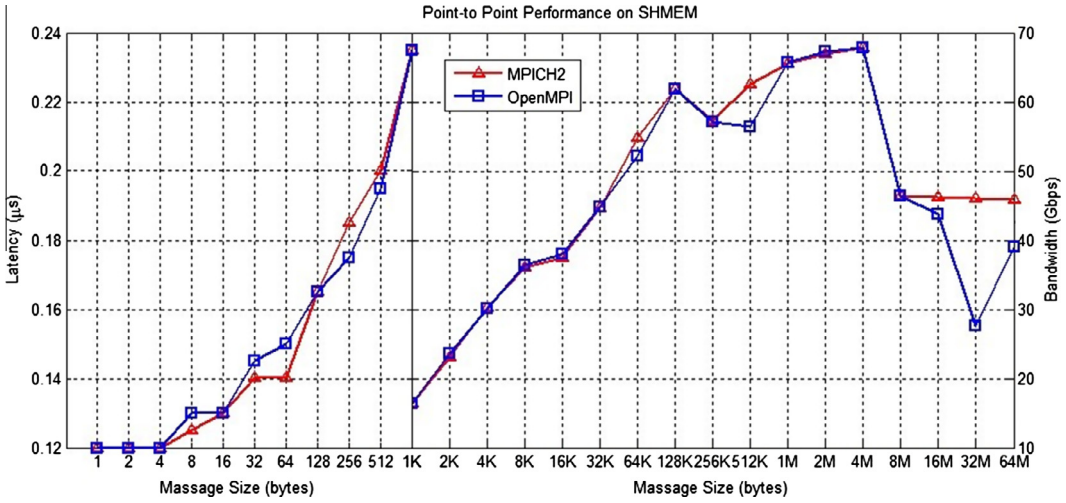

Figure 2 Point-to-point shared memory communication performance on Azure instances.

The point-to point micro-benchmarking results have been obtained with the Intel MPI Benchmark suite (IMB). In addition, NAS Parallel Benchmarks (NPB) kernels, have been assessed using the official NPB-MPI version 3.3.

The evaluation performance metrics of the NPB kernels are MOPS and speedup. MOPS is the Million Operations Per Second of a program, while speedup is the ratio of serial program execution time to parallel program execution time. The larger the value of MOPS and speedup, the better performance we have. A NPB Class C workload has been selected because its performance is highly influenced by efficiency of the communication middleware and the support of the underlying networks.

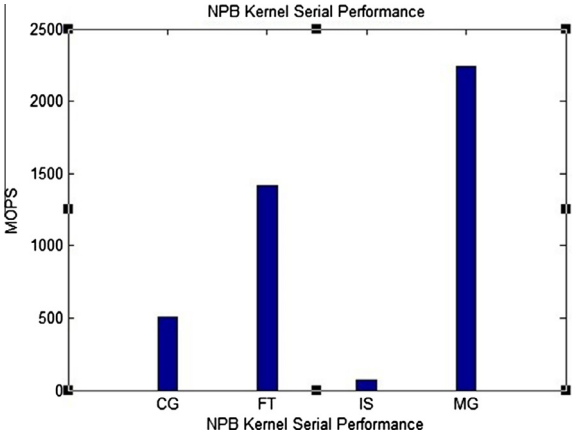

Figure 3 NPB kernels serial performance.

4. Experimental results and discussion

4.1. Inter-VM point-to-point micro-benchmarking

This experiment measures point-to-point latencies for short messages and bandwidths for long messages using both

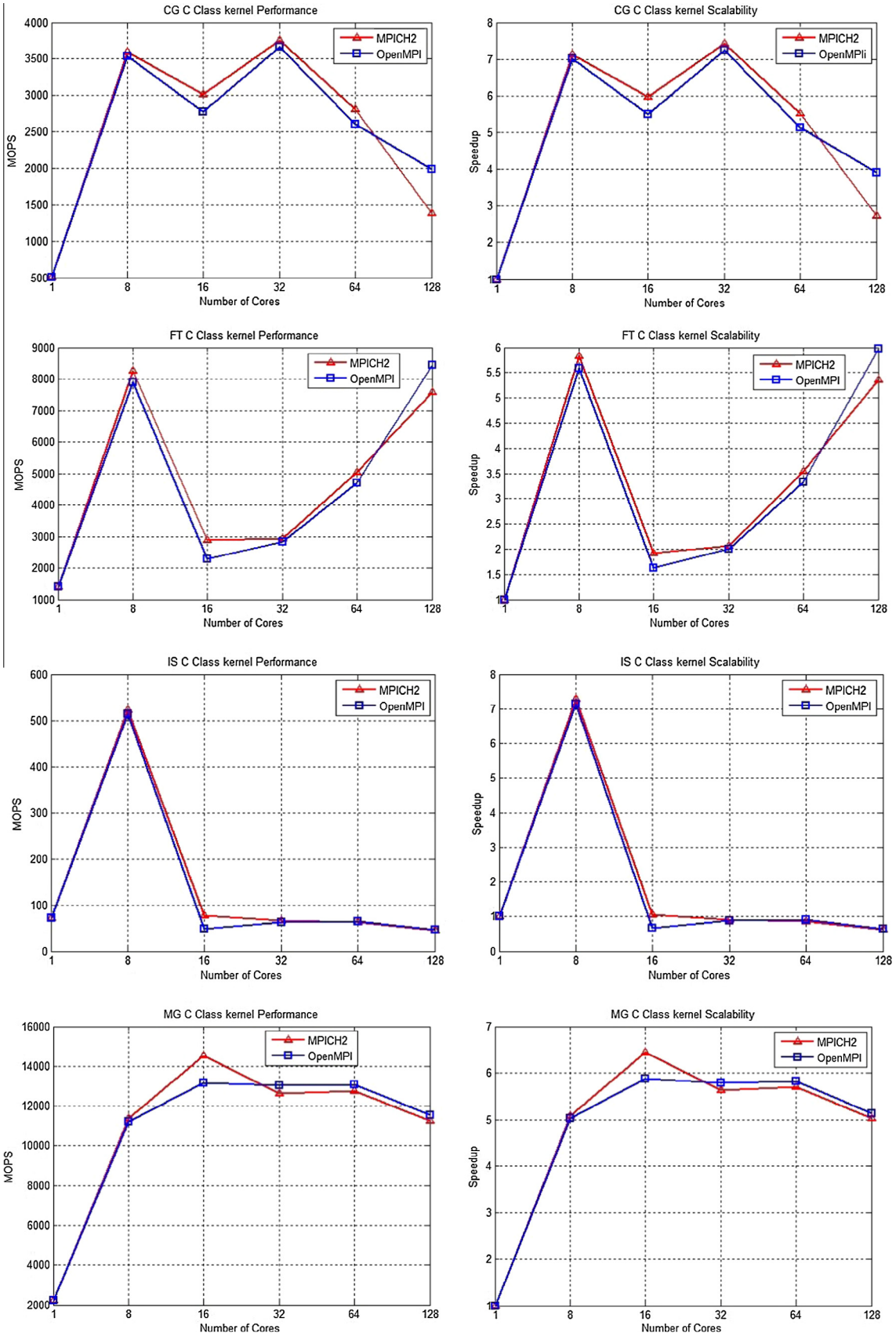

Figure 4 NPB kernels performance and scalability on Azure instances.

MPICH2 and OpenMPI as different Implementations of MPI. An Ethernet scenario communication is performed through an Ethernet network link. Fig. 1 shows the half of the round-trip time of a ping-pong test (in μsec) and the bandwidth (in Gbps) as well. The lowest start-up latency obtained was 33 μsec by MPICH2 which is better than OpenMPI latency (around 55 μsec). The figure shows that MPICH2 is recommended for message sizes less than 128 bytes while OpenMPI is recommended for larger message sizes. Very close values of bandwidth were reported for both MPICH2 and OpenMPI with a maximum value of 1.36 Gbps.

4.2. Intra-VM point-to-point micro-benchmarking

This experiment measures the latency and bandwidth using shared memory model, by sending messages between two processes on the same VM, i.e. intra-VM scenario. As shown in Fig. 2, the latency and bandwidth were better than their correspondents in an inter-VM scenario (see Fig. 1). Similar performance results were obtained for both of MPICH2 and OpenMPI, where a maximum latency was around 0.24 μsec and a maximum bandwidth was around 68 Gbps.

4.3. HPC kernels performance analysis

Four representative NPB kernels, the most communication-intensive codes of the suite, have been used, namely, CG (Conjugate Gradient), FT (Fourier Transform), IS (Integer Sort) and MG (Multi-Grid), using NPB Class C workloads. The experiment is conducted on different sizes of VMs cluster, starting from 1 to 16 VMs with number of cores ranging from 8 to 128 cores correspondingly. The objective of this experiment is to analyze the performance of the HPC kernels on homogeneous clusters with different sizes (configurations). In each configuration, number of MOPS and corresponding speedup were measured.

4.3.1. NPB kernels serial performance

The NBP kernels codes have been compiled using the GNU, showing their MOPS for the serial version as shown in Fig. 3.

4.3.2. NPB kernel performance and scalability

This experiment measures the performance of CG, FT, IS and MG kernels using up to 128 cores on Azure platform (hence, using a cluster of up to 16 VMs). Fig. 4 shows MOPS (left graphs), and speedups (right graphs) of each kernel against number of cores in the cluster. The number of VMs used in the performance evaluation is the number of cores used divided by the number of cores per instance type (8 cores for A10).

The CG kernel showed a maximum performance (both MPOS and Speedup) at a cluster of 4 VMs (i.e. 32 cores), then a degraded performance after that. This degradation indicates a virtualized network performance bottleneck that could be solved by better NIC card virtualization implementation such

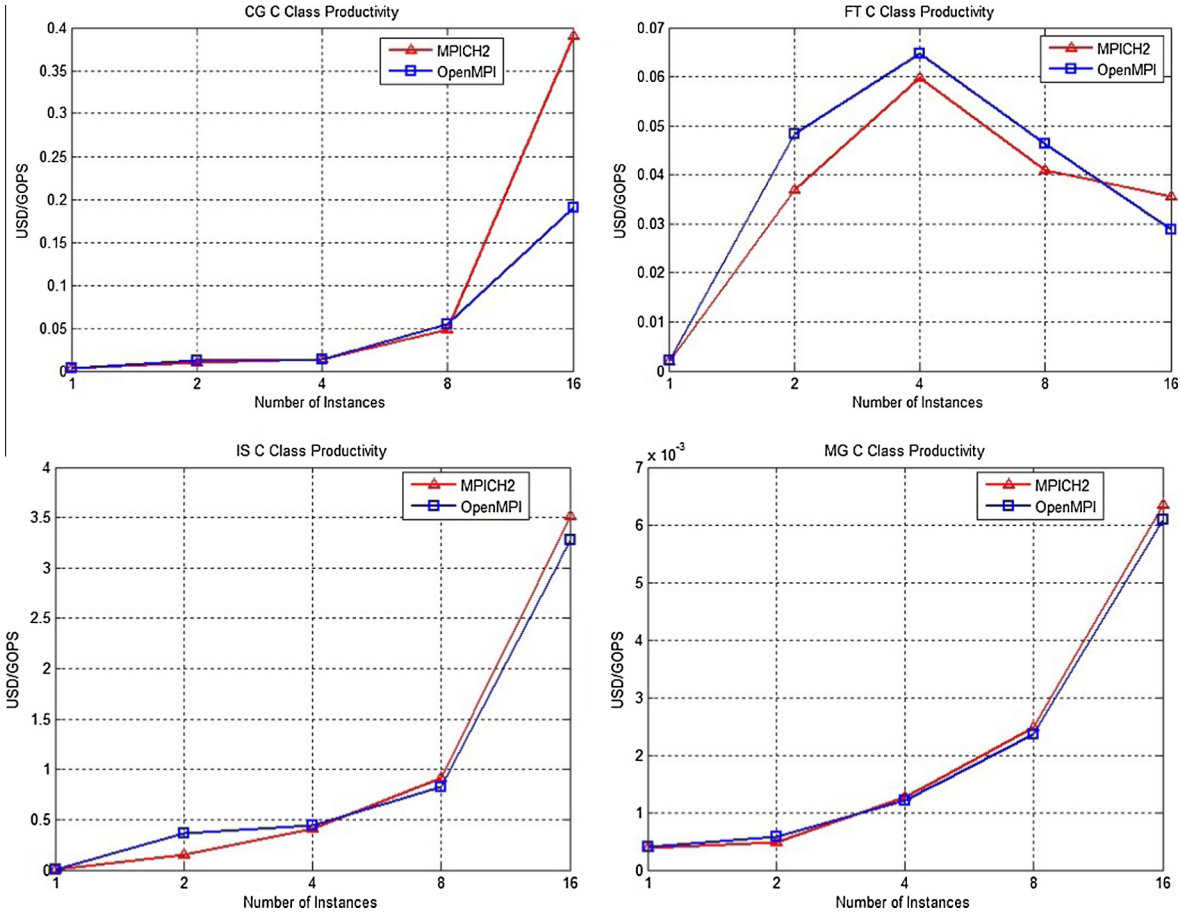

Figure 5 NPB kernels productivity on Azure instances.

as SR-IOV or using faster network fabrics such as InfiniBand. An optimum performance of IS kernel was achieved at clusters of one VM. FT kernels showed similar maximum performance at a cluster of 1 and 16 VMs, under MPICH2 AND OpenMPI respectively.

The analysis of the NPB kernels performance shows that the evaluated libraries obtain good results when running entirely on shared memory (on a single VM) using up to 8 cores in A10 VMs, respectively, due to the higher performance and scalability of intra-VM shared memory communications. However, when using more than one VM, the evaluated kernels scale poorly, experiencing important performance drawbacks due to the network virtualization overhead. The poorest scalability has been obtained by IS kernel where the highest performance was achieved at 8 cores (i.e. one VM). On the other hand, the most scalable kernel was MG, where it achieved an optimum performance at a cluster of 8 VMs (128 cores).

CG kernel, characterized by multiple point-to-point data movements, achieves on A10 its highest speedup value of 7.3 using 32 cores, dropping dramatically its performance from that point on as it has to rely on Ethernet communications, where the network virtualization overhead represents the main performance bottleneck.

FT kernel showed a limited scalability under MPICH2, but a good scalability under OpenMPI. It achieved its highest speedup value of approximately 5.6 using MPICH2 at 8 cores and OpenMPI 128 cores.

IS kernel is a communication-intensive code whose scalability greatly depends on point-to-point communication start-up latency. Thus, this code only scales when using a single VM because of the high performance of shared memory transfers, whereas it suffers a significant slowdown when using more than one VM.

Finally, MG kernel is the less communication-intensive code under evaluation and for this reason it presents the highest scalability on Azure A10 VMs, achieving a maximum speedup value of 10 for OpenMPI.

4.3.3. Cost analysis

The cost of running each kernel on a certain cluster configuration is calculated as USD per GOPS (Giga Operations per Second). Fig. 5 presents the USD per GOPS against number of instances of the cluster. We based our computation on the flat rate of A10 as taken from Azure website.

5. Conclusion and future work

Microsoft Azure Cloud provides the possibility of computing on virtual parallel clusters. Several studies investigated the benefits of performing HPC applications on the Azure Cloud. IMB and NAS parallel benchmarks are used to assess network performance and scalability of HPC application on Azure Cloud. Our results revealed that better performance was delivered by IS kernel and FT kernel (under MPICH2) when running entirely on a single VM (shared memory communication model). Because of the higher performance and scalability of intra-VM shared memory communications, when using more than one VM, the evaluated kernels scale poorly. We plan to extend the study of communication performance and scalability of desired HPC application with better underlying physical interconnection such as InfiniBand. Moreover, a large size cluster (up to 512 cores) as well as different type of Azure instance (with 16 cores) will be studied in the future.

Acknowledgments

The authors express their gratitude for the partial support received by Microsoft to access Azure Cloud computing facilities. We also would like to thank Aya Ibrahim and Mona Kashkoush, from the Informatics Research Institute for their collaboration in conducting the experiments.

References

[1] Gupta A, Milojicic D. Evaluation of hpc applications on cloud. In: 2011 sixth open cirrus summit (OCS). IEEE; 2011. p. 22–6.
[2] Hu Y, Long X, Zhang J. Enhance virtualized HPC system based on I/O behavior perception and asymmetric scheduling. In: 2012 IEEE 14th international conference on high performance computing and communication & 2012 IEEE 9th international conference on embedded software and systems (HPCC-ICESS). IEEE; 2012. p. 169–78.
[3] Sultan N. Discovering the potential of cloud computing in accelerating the search for curing serious illnesses. Int J Inf Manage 2014;34(2):221–5.
[4] Ismail R, Hamid NAWA, Othman M, Latip R, Sanwani MA. Point-to-point communication on gigabit ethernet and Infiniband networks. In: Informatics engineering and information science. Springer; 2011. p. 369–82.
[5] Top500 supercomputing system. <http://www.top500.org>.
[6] Mauch V, Kunze M, Hillenbrand M. High performance cloud computing. Future Gener Comput Syst 2013;29(6):1408–16.
[7] Jackson KR, Ramakrishnan L, Muriki K, Canon S, Cholia S, Shalf J, et al. Performance analysis of high performance computing applications on the amazon web services cloud. In: 2010 IEEE second international conference on cloud computing technology and science (CloudCom). IEEE; 2010. p. 159–68.
[8] NPB: NAS Parallel Benchmarks. <http://www.nas.nasa.gov/Resources/Software/npb.html>.
[9] SKaMPI. <http://liinwww.ira.uka.de/skampi/>.
[10] Mpptest. <http://www.mcs.anl.gov/research/projects/mpi/mpptest/>.
[11] Pallas MPI Benchmark. <https://software.intel.com/en-us/articles/intel-mpi-benchmarks>.
[12] MPBench. <http://icl.cs.utk.edu/projects/llcbench/mpbench.html>.
[13] MPIBench. <http://www.dhpc.edelaide.edu.au/projects/MPIBench>.
[14] HPL – a portable implementation of the highperformance linpack benchmark for distributed-memory computers. <http://www.netlib.org/benchmark/hpl/>.
[15] ExpóSito RR, Taboada GL, Ramos S, TouriñO J, Doallo R. Performance analysis of HPC applications in the cloud. Future Gener Comput Syst 2013;29(1):218–29.
[16] Jose J, Li M, Lu X, Kandalla KC, Arnold MD, Panda DK. SR-IOV support for virtualization on infiniband clusters: early experience. In: 2013 13th IEEE/ACM international symposium on cluster, cloud and grid computing (CCGrid). IEEE; 2013. p. 385–92.
[17] Infiniband Trade Association. <http://www.infinibandta.org>.
[18] Roloff E, Diener M, Carissimi A, Navaux POA. High performance computing in the cloud: deployment, performance and cost efficiency. In: 2012 IEEE 4th international conference on cloud computing technology and science (CloudCom). IEEE; 2012. p. 371–8.
[19] Akioka S, Muraoka Y. HPC benchmarks on Amazon EC2. In: 2010 IEEE 24th international conference on advanced information networking and applications workshops (WAINA). IEEE; 2010. p. 1029–34.

[20] Hassani R, Aiatullah M, Luksch P. Improving HPC application performance in public cloud. IERI Procedia 2014;10:169–76.
[21] Belgacem MB, Chopard B. A hybrid HPC/cloud distributed infrastructure: coupling EC2 cloud resources with HPC clusters to run large tightly coupled multiscale applications. Future Gener Comput Syst 2015;42:11–21.
[22] Evangelinos C, Hill C. Cloud computing for parallel scientific HPC applications: feasibility of running coupled atmosphere–ocean climate models on Amazons EC2. Ratio 2008;2(2.40):2–34.
[23] Mehrotra P, Djomehri J, Heistand S, Hood R, Jin H, Lazanoff A, et al. Performance evaluation of Amazon EC2 for NASA HPC applications. In: Proceedings of the 3rd workshop on scientific cloud computing date. ACM; 2012. p. 41–50.
[24] Azure compute size. <http://azure.microsoft.com/blog/2015/03/05/new-a10a11-azure-compute-sizes/>.
[25] Openmpi website. <http://www.open-mpi.org/>.
[26] Mpich2 website. <http://www.mcs.anl.gov/research/projects/mpich2.org/>.