Computing (2013) 95:89–108
DOI 10.1007/s00607-012-0213-0

Performance issues and performance analysis tools for HPC cloud applications: a survey

Shajulin Benedict

Received: 12 January 2012 / Accepted: 16 August 2012 / Published online: 9 September 2012

Abstract Cloud Computing is an eminent emerging technology that surpasses Grids from their IT resource administrations and arduous Grid middleware solutions. At present, users could access an abundant number of pre-defined cloud services or run their programs on demand as a pay-as-you-go processing model without much distribution problems. In addition, the IT business market has pumped enough revenue for establishing salient common-use cloud solutions. Despite adequate researchers have been involved in the cloud development, scientific application developers are still reluctant to execute their applications in the cloud due to the performance concerns, such as, scalability, availability, and service level agreement violations of the cloud providers. In this paper, a survey of various High Performance Computing (HPC) applications and possible performance concerns while executing applications in cloud is presented. Pointing out the need for Performance Analysis (PA) tools, this paper focuses on the study of cloud-based PA tools in detail. This paper could leverage HPC application developers to cope with the performance issues and to best utilize the available performance analysis tools of clouds.

Keywords Cloud applications · Cloud computing · HPC · Performance analysis tools · Service level agreements

This work is partially funded by the HPCCLoud project, an ongoing research grant, under Returning-Experts programme of CIMOnline, GIZ, Germany, and Department of Science and Technology, India.

S. Benedict (✉)
HPCCLoud Reserach Laboratory, St.Xavier's Catholic College of Engineering, Anna University, Nagercoil 629003, India
e-mail: shajubenedict@yahoo.com

S. Benedict
e-mail: shajulin@sxcce.edu.in

Mathematics Subject Classification 68-02 · 68Q25 · 68W40

1 Introduction

HPC is a long standing research keyword stemming from the scientific application developers community. Earlier, application developers were using supercomputers or mainframes with single administration, limited sharing, and less user control to solve tedious computational problems. Not all application developers were allowed to access those machines due to the stricter policy, privacy, and cost issues. However, with the advancement in IT technologies, such as, cluster computing or grid computing, the perspective of solving HPC applications advanced.

Recently, cloud computing, with its dynamically scalable virtualization technique and a cost effective pay-as-you-go model of IT resource sharing, has motivated HPC application developers to utilize cloud for solving their applications, even those requiring exa-scale computations, without much difficulty. With cloud's inordinate features, cloud applications have emerged tremendously in diverse fields, such as, business, social-networking, scientific, enterprise, and content delivery domains. For instance, some commercial HPC applications available in the market include, simulation of car crashes, new drug designs, and airflow over automobiles or airplanes. More important aspect is that the engineering companies of HPC domain and a few small computational *science and engineering research groups* [20] are planning to maximize their profits and clients using cloud environments.

To highlight cloud computing, some fundamental viewpoints—definition, classification, and importance—are expressed as below:

- *Definition*: According to Ian Foster, cloud is a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet [22]. Mell and Grance [35] defines cloud as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.
- *Classification*: The cloud service models [35] are classified as Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as Service (IaaS). The deployment models of clouds are classified as private, public, community and hybrid models.
- *Advantages*: Clouds excel clusters or grids [25] in five different aspects: (i) maximizing profits using online support, (ii) deploying with reduced manpower, (iii) providing abundant solutions for any scientific or business problems, (iv) minimizing the complexity among resource connectivity, and (v) increasing competence and productivity.

Although several HPC initiatives have started utilizing cloud environments, performance concerns of HPC applications in cloud are immense. For instances, HPC applications could suffer from load imbalance, scalability, data management, and security concerns. The most well known performance issues that arise in data-intensive HPC

Table 1 Performance analysis tools—role, capabilities, and challenges

Role of PA tools	Capabilities of existing PA tools	Challenges of existing PA tools
(i) To understand performance problems	(i) Finding the performance of cloud—VM creation, overload, idleness, VM instances, and so forth	(i) To find the performance of applications using h/w counters from VMs
(ii) To express performance problems in a user friendly manner	(ii) Expressing output as html, xml, cmd line or as more intuitive approaches	(ii) Providing comprehensive and meaningful findings of performance issues
(iii) To scale well with less overhead	(iii) Immediate execution for PA without waiting in submission queues as in traditional HPCs	(iii) Huge overhead because of the VM creation and migration in virtualization approaches, such as, hypervisors

cloud applications are memory management issues, data locality, data mobility, data storage, legal aspects, and security vulnerability issues. Similarly, the performance issues of compute-intensive HPC applications emerge while mapping and reallocating jobs to the dynamic cloud resources, mostly virtual machines. It is observed in [29] that the disk I/O, latencies of process create, and network communication overheads in multiple virtual machine system are the most crucial factors that challenge the performance of clouds. These performance concerns should be notified to the end user or application developer or cloud provider as required by them at the earliest.

In this context, in recent years, emerging cloud-based performance analysis tools aim at playing a vital role for users or cloud providers. The most important roles of such performance analysis tools and their capabilities and challenges are highlighted in Table 1. As it can be seen, PA tools could suggest users to find performance problems of applications that are related to Virtual Machines (VMs) or idleness, immediately. However, the challenge of obtaining performance measurements from hardware counters of VMs and overhead due to virtualization remain as a valid point for near future research considerations.

Some researchers have studied the performance of specific HPC applications on cloud [1]. A few studies have compared the performance of cloud providers [4,16] with sample cloud applications. However, there exists a very few research works that explores the performance concerns of HPC cloud applications and existing cloud-based performance analysis tools although various efforts have been accomplished through developing performance monitoring tools for HPC applications.

The main contributions of this paper are as follows:

1. Surveying the existing works on HPC cloud applications.
2. Exploring the possible performance concerns of HPC cloud applications.
3. Expressing the need for performance analysis tools and their challenges.
4. Surveying the existing performance analysis tools that address the performance concerns of HPC cloud applications.

The rest of the paper is organized as follows. Section 2 presents existing cloud applications and benchmarks. Section 3 explains the possible performance concerns

of HPC cloud applications and Sect. 4 discusses the performance analysis tools for cloud applications. Finally, Sect. 5 presents a few conclusions.

2 Cloud applications

Traditionally, cloud applications were widely deployed for executing commercial web applications. Cloud applications were classified as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), or Software as a Service (SaaS) using three deployment models, namely, Public, Private, or Hybrid cloud instantiations. Recently, Tobias et al. [21] has classified Cloud applications from the perspective of providing quality services to end users as business, personal, multimedia-intensive, gaming, and so forth.

Cloud technology was quite successful in terms of scalability and availability for users who were not much concerned about performance issues. It had rooted its strong base for academic support which includes interactive classes, online tutorials, lab facilities, compute resource provisioning, and so forth. Similarly, the technology had enhanced support over rare traditional courses which could not be opted in hostile environments due to non-availability of resource persons. For instance, Embedded programming [11] was hosted in cloud for academia and Indian Carnatic music is hosted in cloud for public.

However, in recent years, a stream of HPC application developers [34] has seriously thought about utilizing cloud technology [8] due to its flexible and cost effective approach to accessing compute nodes by leveraging various virtualization technologies, including hardware assisted virtualization. Gideon Juve et al. [27] has expressed the importance of cloud for solving HPC applications by analyzing the relationship between scientific workflows and clouds. In addition, the alternative approach of handling issues while framing a virtualized HPC cluster, which was proposed by Georg et al. [9], has promoted HPC application developers to opt cloud infrastructures for solving HPC applications. This kind of support to cloud establishments or allied promotional technologies (VMs) has led the researchers to study the feasibility and performance concerns of HPC applications on cloud.

These performance studies were mooted up in a large-scale due to the voluminous support obtained via funding agencies, such as, National Science Foundation [26], Department of Science and Technology [45], Grid 5000 project, and National Bioinformatics Network. Similarly, many international industry initiatives, such as, Climate Savers Computing Initiative (CSCI), Green Computing Impact Organization (GCIO), Green Electronics Council, The Green Grid, International Professional Practice Partnership (IP3), jointly with some leading companies (IBM, Amazon, Google, Intel, and HP) [5] have projected cloud via end users utility [38]. This section explains the wide utility of cloud for solving HPC applications and the respective domain areas.

2.1 Existing HPC cloud applications

HPC-based applications have endeavored utilizing clouds from various disciplines including seismic, bio-technology, drug design, and so forth. A few domain areas

where cloud is widely being used, the corresponding cloud applications, and possible performance problems (see Sect. 3) are listed as below:

- *High Energy Physics Domain*: High Energy Physics-based applications are, in general, data-intensive where the application receives large sets of input data at higher rates. Charbonneau A. [48] has studied the fastest streaming of large data sets from single location to each clouds while experimenting *BaBar application* that recorded electron-positron collisions at the SLAC National Accelerator Laboratory from 2008 to 2009. Similarly, modern high-energy physics experiments, such as DZero1, typically generate more than one TeraByte of data per day [46].
- *Geographic / Seismic Domain*: Few initiations are carried out in HPC community to solving geographic / seismic applications, such as, climate change prediction, weather prediction, seismic analysis, and so forth, using cloud environments. Such applications have aimed at addressing cloud related performance issues, namely, managing large data sets, accessibility, and scalability. Most of the Geographic / Seismic applications are data-intensive and sensitive to memory management issues (to be discussed in next section).
- *Electronics Design Community*: Various HPC solutions from Electronics Design community, such as, *Static Timing Analysis*, Computational Lithography for process modeling, predictive models of designs, and Simulation study of PCB design, are planning to utilize cloud resources in order to reduce cost and speed-up the production.
- *Bioinformatics Domain*: Bioinformatics, in general, has played a vital role in HPC community. A few researchers have ventured in studying the possibility of using cloud technology. For instances, Accenture and ATandT have launched a cloud-based medical imaging service that will help health professionals and radiologists access, review, and store X-rays, CT, MRI scans, and other images through ATandT's network [12]. Metabolomics application which studies the chemical processes of metabolites was experimented using SCALE. SCALE is known as an analysis platform for e-Science experimental data [42]. Most of the applications from Bioinformatics domain suffer from performance issues, namely, memory management and poor response time.
- *Media and Gaming Domains*: Cloud computing for media [15] and entertainment industries is growing due to a low cost requirements for the massive storage and retrieval of data. To illustrate with an example, a small post-production studio with two editing systems may need TerraBytes of data storage requirements. This is usually done using discs, DVD or VCD, and then transported to some other storage devices. On contrary, if there is a cloud environment with a high-end network connection, the data could be uploaded, worked, or accessed using IP addresses. This avoids the need for a massive storage resource requirements at the studio. Similarly, game industry [6] has indulged its proficiency as competing human brains while solving gaming problems using the artificial intelligence and distributed compute resources - cloud has added up the support through its enhanced scalability features. Although the media and gaming domain found cloud as advantageous, the resource management and provisioning issues, currently, hinder their wide utility.
- *Large-scale Engineering Simulation Studies*: Automobile and Aeronautical industries have started experiencing the cloud footprints. Most of the simulation studies

and predictive modeling before production are solved in a distributed manner using clouds. For instance, the IBM Engineering Solutions for Cloud can provide a blueprint and a foundation to quickly set up and manage highly efficient and secure electronic design automation and computer aided design and analysis clouds to streamline engineering processes, reduce costs, and design cycle times. Applications, such as, structural design, concrete modeling, automatic building plan, design analysis, and so forth are deployed as cloud services for civil engineers or building constructors. In general, researchers working with large scale engineering simulations have little knowledge about hardware architectures and performance concerns. They rely on performance analysis tools, if any, for enhancing the performance of their codes.

Apart from running real-world HPC applications in cloud, a few researchers have studied the feasibility or performance issues of clouds using some standard available benchmarks. Most notable benchmarks that were used for evaluating the performance of clouds include High Performance Linpack (HPL) benchmark of High Performance Computing Challenge, NAS Parallel benchmarks, NERSC benchmarks [41], and so forth. Knight et al. [31] has evaluated the efficacy of cloud for cluster computations using benchmarks in terms of internet throughput, latency, CPU cache throughput, RAM throughput, Inter-process latency, and so forth.

Although there exist wide utilities of HPC applications in clouds, a few researchers, perhaps, have argued from their recent evaluation-based studies that cloud technology will be beneficial only for selective HPC applications due to underlying poor network performances [19] and the other performance issues (see Sect. 3).

3 Performance concerns

HPC applications endeavor performance challenges in various stages of their execution on cloud. This section discusses the possible performance issues of HPC cloud applications when executed in private, public, or hybrid clouds.

Performance is generally tied to the application's capabilities within the alien cloud infrastructures [43]. Developing and deploying those applications on cloud are not so arduous than serving them to the users as asserted. Persuading reasons that lead to a reduced performance while running HPC applications on cloud are illustrated in Fig. 1.

3.1 Resource provisioning issues

Running HPC applications efficiently in private, public, or hybrid clouds remains challenging due to poor memory management among VMs, poor response time, dynamism on cloud resources, or improper resource prediction mechanisms.

3.1.1 Memory management issues

Mostly cloud uses hypervisor technology, a technology that has hardware virtualization using Virtual Machine Manager (VMM) which allows multiple guest operating

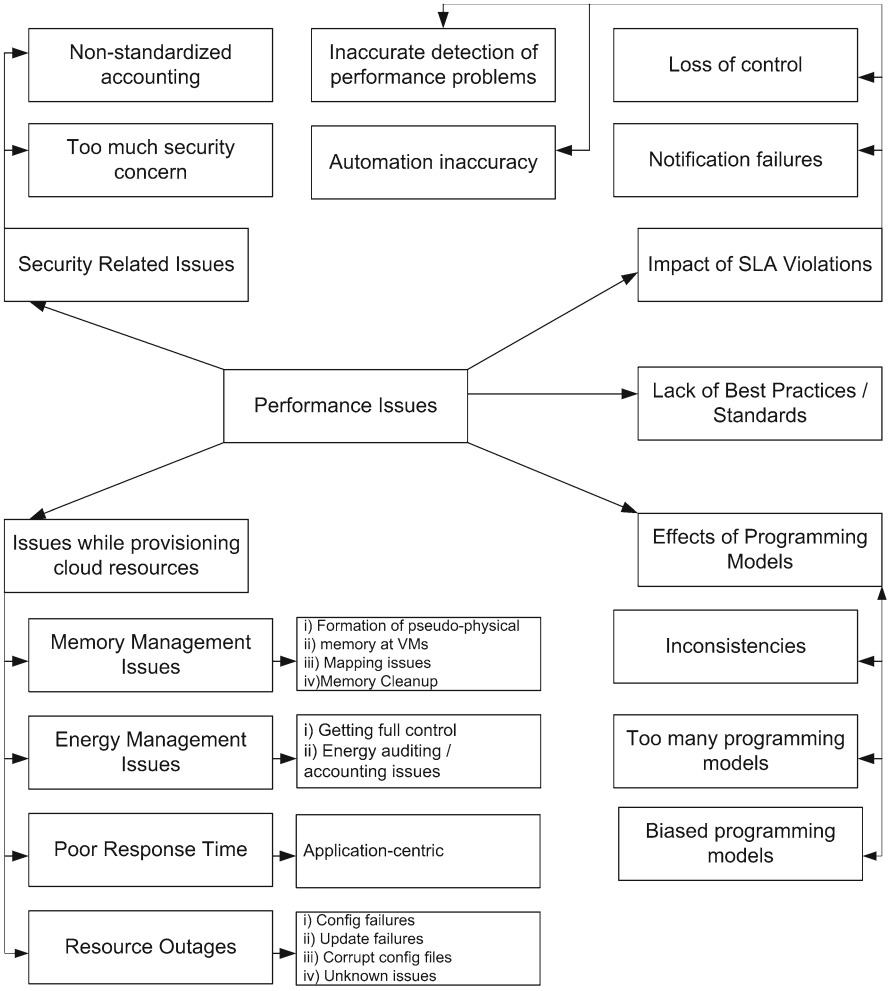

Fig. 1 Possible performance issues of HPC cloud applications

systems to run concurrently on a host computer. Those multiple operating systems receive a pseudo-physical memory—contiguous range of physical page frames starting at physical frame 0, despite the fact that the underlying machine page frames may be sparsely allocated and in any order—where the memory cannot be guaranteed. The Xen Hypervisor, for instance, maintains VM specific table with the mapping of pseudo-physical memory to real memory and a global mapping table that records the mapping from machine page frames to pseudo-physical memory.

Formation of pseudo-physical memory, remapping them to actual machine page frames, and memory cleanup process using VMM might lead to huge overheads, delays, or memory leakages. If the cleanup process fails, the memory frames alloted for a VM could not be reused by new ones. This degrades the performance due to hefty memory management overhead and resource under-utilization.

3.1.2 Energy management issues

Energy management is a serious performance concern for virtualized cloud resources due to the non-awareness of emerging guest OSs. This can cause an OS to lack full control over the hardware resources for providing an efficient energy management mechanisms [5]. In addition, the energy auditing or accounting for guest OSs will be error prone when the hardware resource is shared by multiple guest OSs. In such cases, the system needs to invoke energy-efficient policies challenging the completion time of applications. The energy management issues are more challenging for cloud-based mobile applications [28].

3.1.3 Poor response time

Response time is more centric to applications. Enterprise applications in cloud endeavor poor response time when users run applications from remote sites where the latency of WAN or Internet link between the data center, cloud location, and the user is high. Scientific applications can lead to poor response time due to failed checkpointing or non-availability of cloud resources.

3.1.4 Resource outages

Resource outages are unavoidable which heavily challenges the performance of clouds. Cloud outages might occur due to configuration failures, update failures, corrupt configuration files, technically challenged expert operations, or even unknown issues. This could lead to serious business disruption or unexpected loses. For instance, Microsoft's cloud services had a 3 hour long resource outage when an updating process was undergone [36].

The other aspect would be due to a failure to boot up a cloud instance which is hosted on a hybrid cloud implementation, especially after VM migration or loading device drivers happened [7].

3.2 Effects of programming models

There exists various leading programming models for clouds, such as, task model, thread, MapReduce, PSM, Workflow, MPI, and Actors. Most of these programming models have inconsistencies between the fault recovery mechanisms in execution and storage processes. Inconsistencies can lead to broken service instances and hence a poor performance. For instance, the MapReduce architecture and programming model pioneered by Google is an example of a modern systems architecture designed for processing and analyzing large datasets and is being used successfully by Google in many applications to process massive amounts of raw Web data. This programming model has no mechanism for fault recovery procedures.

3.3 Impact of SLA violations

Service terms of clouds are generally expressed in SLAs which benefits both cloud providers and consumers. However, the existing service level management mechanism

which includes an infrastructure for automatically negotiating, creating, managing, and enforcing terms of SLA faces performance challenges due to loss of control over resource provisioning or deploying cloud services in cloud. Additionally, if the SLA violation is not notified at the initial stage, the defaulters would have to pay huge penalties. This urges the need for a standard practice for handling SLA violations so that the applications are solved within a limited time frame [44].

3.4 Lack of best practices or standards

Lack of best practices or standards in clouds has severely prevented the betterment of cloud from its utilization [10,47]. By this, the users are urged to select some industrial compute resources without competence; the users might have to write redundantly their applications with slight modifications to different providers if there is no standard interoperable cloud interfaces. Additionally, the lack of standards would have impacts on developing transparent code. Although there exist some proposals for cloud standards which emerged from open source community, they are not practiced.

3.5 Security issues

Cloud is still a perilous technology for security concerned organizations, especially, financial dealers. Although unavoidable, clouds prefer more sophisticated security measures challenging performance aspects. For instance, providing data security thorough encryption and decryption mechanisms [18] do have performance impact, namely, reduced response time. Similarly, when there is a requirement of enhanced mechanisms such as identity management, data locality management, and so forth, the performance is affected. Hence, increased security would have a direct impact on the performance of cloud applications.

Another important issue that drags down the performance of clouds is a non-standardized accounting. To illustrate, let us consider the existing market models for pricing cloud applications. It could be noticed that the billing options are different for various cloud providers, Amazon EC2, Google AppEngine, or Microsoft Azure, which can lead to fraudulent billing or inefficient pricing [49].

In addition, policy or legal issues obstruct the smooth functioning among cloud providers, thereby leading to performance issues. For example, executing games on a cloud server can promote illegal copying of those applications by deceptive cloud providers which creates a tug-of-war on further provisioning of cloud resources. A detailed study of security issues on cloud can be seen in [33].

4 Performance analysis tools

In order to pinpoint performance bottlenecks in clouds (see Sect. 3), there is a need for performance analysis tools. Existing Performance Analysis (PA) tools help users in writing HPC applications on parallel machines. These tools can provide the user with measurements of the programs performance and pinpoint locations for bottlenecks.

Table 2 Comparison of performance analysis tools

Tool	Category	Owner	Monitors	Frequency	OSSupport	Online	Usability	Output	Open source
ACW	API	Amazon	AWS Cloud resources such as EC2, RDS DB instances	Every 5 min (free) and 1 min (charge)	Any	Online	Inbuilt	Graphs, alarms	No
Ganglia	API	Univ. of California	CPU, memory, disk, network, and number of VMs running	Every 30 s	Ubuntu. Otherwise, additional package should be installed	*	Installation specific to different clouds	Web visual-ization	Yes
Nagios	Plugin	Nagios Inc.	Network outages, CPU, memory, disk details, availability checking	Can be configured	Ubuntu. Otherwise, to configure	*	Need to install	Views, reports, emails, alerts	Both
Azure	Appl.	Microsoft	Health status of servers, applications, and clients	Every 5 min	Application	Online	Easy	Graphs	No
InterMapper	Plugin	Intermapper Inc.	Focuses on network performance metrics	As in ACW	As in ACW	As in ACW	As in ACW	As in ACW	No
LogicMonitor	Agents	LogicMonitor Inc.	Monitors newly added or deleted cloud instances	*	*	Online	*	Views, alerts, emails	No
CloudStatus	Plugin	Hyperic Inc.	More metrics are monitored including health status	Continuous	Any	Online or weekly	Simple	Reports	Both

Table 2 continued

Tool	Category	Owner	Monitors	Frequency	OSSupport	Online	Usability	Output	Open source
CloudMonitor	API	Nimsoft Inc.	Server performance, resource usage	*	Any	Online	Simple	Alarms	No
CloudKick	Plugin	CloudKick Inc.	Server performance	Automatic	Any	*	Simple via deployment tool integration	3D visualization, reports	No
InfoManager	Sensor-based	OpenNebula	Host details, usage, available CPU	*	Ubuntu	online	Too much configuration	Report	Yes

* Information is either not applicable or unknown

There exists some well known PA tools, namely, Paradyn, Periscope, TAU, Vampir, KOJAK, SCALASCA, and mpiP for HPC machines with some concrete realizations. However, only a few tools exists for analyzing performance problems on clouds. This section explains the need for PA tools and existing performance analysis methodologies (Table 2) in clouds.

4.1 Need for PA tools in clouds

PA tool is mandatory for HPC cloud users due to the following reasons:

1. To understand memory management problems—pipeline stalls, page updation, memory cleanup failures, and so forth.
2. To report the user about the resource outages and expected vulnerability period of outage.
3. To inform the user the performance problems caused due to the programming models used. For example, OpenMP programming model could create 'n' threads and use more serial constructs which would lead to resource underutilization. Similarly, the MPI programming model might cause latency in communication within cloud processes.
4. To identify the lack of well defined SLAs by cloud providers. Fluctuations in any means shall be notified to the user at the earliest. For instance, researchers in Australia conducted stress tests to demonstrate that Amazon, Google, and Microsoft suffered from variations in performance and availability due to loads. Specifically, the researchers measured how the cloud providers scaled up and responded to the sudden demand of 2,000 concurrent users. In some cases of their study, response times at different points of the day varied by a factor of 20 [2].
5. To be aware of the providers who might allocate the resources that are energy inefficient, non-scalable, Service Level Agreement (SLA) violated, and those that are less secure. Cloud providers are not unique. The analysis or optimization strategy reports would help them to select the providers. To illustrate, some cloud providers has increased data storage performance with increased load, whereas, it is vice versa for the others. As a note, neither Google App Engine nor MS Azure, scale linearly like electricity.
6. To receive the best service based on their requisition, for example, by mapping appropriate virtual machines to servers.
7. To contribute in reducing the environmental impact due to carbon dioxide emissions while powering on those energy inefficient cloud resources.

4.2 Available PA tools for clouds

4.2.1 Amazon cloud watch

Amazon CloudWatch [3] is a web service based PA tool designed by Amazon Inc. that monitors cloud resources, such as, Amazon EC2, Amazon RDS DB instances, or application centric metrics both at free of cost or additional charges. It enables users to collect, view, and analyze pre-defined metrics after following five steps as follows:

1. *Signup* defaults any Amazon services. Amazon cloud watch also authorizes the user using signup mechanism.
2. *Setting up Command Line Interface* (*CLI*) is the second step to utilizing the tool. The Command Line Tool serves as the client interface to the Amazon CloudWatch web service. Once after downloading the client interface, the user needs to setup environment variables, add metrics, and configure alarm actions based on data from metrics in the respective files.
3. Then, using CLI, the user can *publish metrics* for monitoring them.
4. *Receive metrics* based on the published metrics using a command mon-get-stats on CLI. The statistics can be average, sum, or relative values.
5. Finally, the obtained statistics of the monitoring results are depicted in a graphical views.

The drawbacks of the tool are as follows: (i) it supports only a few metrics and (ii) the tool is focused on monitoring Amazon cloud resources.

4.2.2 Ganglia

Ganglia is a open-source scalable monitoring tool developed by the University of California for clusters or grids. Satisfying the cloud requirements, namely, scalability and support for monitoring virtual pool of resources, Ganglia extended to clouds such as Eucalyptus and Rackspace. Ganglia remains a robust tool for monitoring server performance metrics using sFlow agents. The important features of Ganglia include distributed configurations, performance metrics for monitoring the number of active VMs, hierarchical structure of monitoring mechanisms, and web-based visualization of results.

Ganglia consists of gmond daemons (agents), gmetad, and a web frontend components. It achieves scalability using local gmond instances to push the data to more central nodes in a hierarchical fashion. The *gmetad* collects data about a collection of resources by periodically polling gmond instances to retrieve the monitoring data and storing the monitoring data in a round-robin database. The web frontend component is user-friendly and it is responsible for collecting the performance data and displaying the results in the form of timeline charts [17].

The installation of Ganglia is much easier with Ubuntu. Otherwise, the user needs to install some additional packages to use the tool. Ganglia is utilized widely by the Eucalyptus and Rackspace cloud community.

Some known issues of Ganglia are that (i) the gmond daemons need to know the ip addresses of their hierarchical daemons as Ganglia works with multicast networking, (ii) Possibility of huge communication overhead while processing the gmond daemon identities, and iii) different configurations are required at higher levels of gmond daemons.

4.2.3 Nagios

Nagios is more related to Ganglia for technical competence when considering cloud monitoring solutions. Nagios is capable of monitoring a variety of servers and

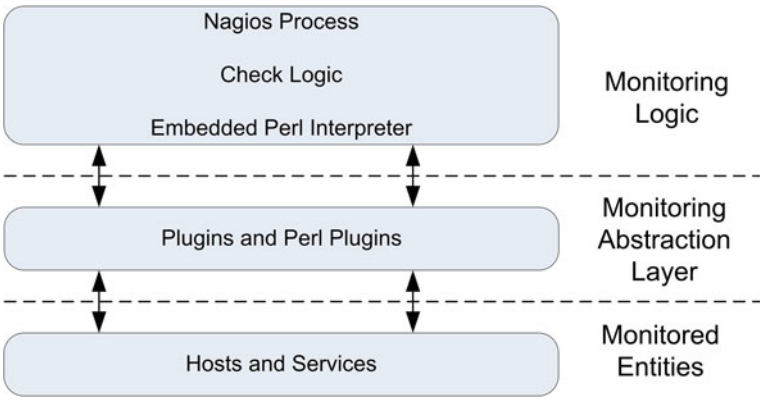

Fig. 2 Architecture of nagios performance analysis tool [37]

operating systems—both physical and virtual—with industry standard. This tool is widely used by many organizations including companies.

Nagios has two cloud monitoring solutions, namely, Nagios XI (commercial) and Nagios Core (open source). The important features of Nagios XI are a PHP web interface, integrated performance graphing, customizable dashboards, web configuration GUI, configuration wizards, user management, and the others. Nagios XI is developed on the top of Nagios Core with additional open source components for improved features [37].

Nagios Core, a open-source monitoring solution, has four important components (i) Nagios Core which contains the core monitoring engine and a basic web interface, (ii) Nagios Plugins which allows the user to monitor services, applications, metrics, and more, (iii) Nagios Frontends which enhance the Nagios experience with additional frontends, and (iv) Nagios Addons which trick out the Nagios installation process with hundreds of addons. Nagios is based on plugins—some compiled executables or scripts (Perl scripts, shell scripts, etc.) that can be run from a command line to check the status or a host or service. Nagios will execute a plugin whenever there is a need to check the status of a service or host and returns back the results to Nagios for showing the results in a prescribed fashion or undergoing some actions via event handlers. The plugin architecture of Nagios is depicted in Fig. 2.

4.2.4 Windows Azure Diagnostic

Windows Azure Diagnostics Monitor is an application with two main components, namely, a website and a scheduler, to view and analyze the cloud performance data in the form of counter graphs, logs, or statistical views. The tool monitors the health states of servers, clients, and applications in an end-to-end fashion for enterprise IT environments.

The diagnostic data are collected independently for processing using a separate storage account. Currently, the tool has three monitoring options, namely, event monitoring, user perspective monitoring, and performance monitoring. Although the tool

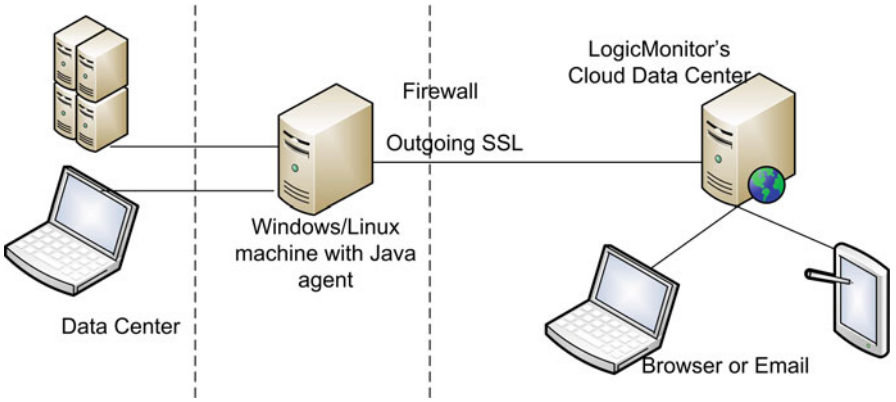

Fig. 3 Architecture of LogicMonitor performance analysis tool [32]

has performance counters, for e.g., *SampleRate and scheduledTransferPeriod*, the performance monitoring option is not yet enabled for users [50].

4.2.5 InterMapper cloud monitor

InterMapper Real-Time Network Knowledge, a company from USA, has developed a plugin jointly with Dartware solutions Inc. for Amazon CloudWatch. This plugin has the capability to monitor network performance metrics for Amazon EC2 Instances and alerts systems administrators to threshold violations. The other benefits of this plugin includes (i) simple deployment for newly launched virtual servers, (ii) process level monitoring and analysis, and (iii) maintenance of historical data to enable root failure analysis [24].

4.2.6 LogicMonitor

LogicMonitor is a monitoring tool that alerts users via email or sms and expresses the monitored data in various forms if paid. The important feature of this tool is simplicity. The tool offers one java agent which should be installed on a server or a machine to be monitored within the organizations firewall. The LogicMonitor's Cloud monitoring center would process send the instant monitoring data in various graphical views at realtime.

The architectural diagram for LogicMonitor is shown in Fig. 3 [32].

4.2.7 CloudStatus

CloudStatus [14] is a plugin from Hyperic Inc. which continuously reports on performance metrics, more than 50,000 metrics, at real time or in weekly basis for web-based applications. Having both open source and enterprise edition, the CloudStatus attracts many researchers mainly to know the service availability, response time, latency, and network health conditions of their applications.

CloudStatus works on the top of Hyperic HQ. Hyperic HQ was designed to monitor and manage large scale web infrastructure. It collects data from various clouds to calculate the overall availability and then normalizes metrics across the cloud. Currently, it has restricted support to Amazon and Google appengine clouds.

4.2.8 Nimsoft cloud monitor

Nimsoft CloudMonitor [40] is designed to support Rackspace cloud customers. It is freely available only for 90 days. It emerged as a result of collaborative effort between Nimsoft and Rackspace. However, Nimsoft has many other server solutions in addition to the CloudMonitor. The main objective of this tool is to track cloud usage, performance, and availability of servers hosted by Rackspace. As common to the other tools, the CloudMonitor has options to set alarms when the monitoring data exceeds some threshold.

4.2.9 CloudKick

CloudKick monitors cloud resources and reports on the performance details as alerts, emails, sms, or 3D visualization. Although it is a paid software, rackspace and some leading cloud solutions have augmented CloudKick monitoring due to its better support for modern custom APIs. With the Cloudkick's simple monitoring scripts, users can monitor any user-defined metrics included to the applications, for instance, the number of users logged in. Cloudkick, additionally, provides REST API, by which the user can query information about servers and integrate Cloudkick into some well-known existing tools and applications [13].

4.2.10 InformationManager

Information Manager (IM) [23] is a monitoring tool developed by an open source OpenNebula cloud community with a vision of bringing an industry standard. IM has used various sensors to identify the performance of servers or cloud resources. Additionally, IM has used separate sensors to collect information from different hypervisors such as Xen or KVM.

The monitoring approach is not so user-friendly due to the requirement of configuring each sensors separately and the requirement of IM drivers. However, the tool can be easily downloaded and updated by Ubuntu.

4.2.11 Supportive tools

There exists various other tools which could be a supportive candidate for cloud monitoring. For instances:

– Munin—Munin is a monitoring tool used for analyzing the CPU usage and network usage. With its graphical rich framework, it has the capability to produce the performance reports user-friendly. The features of this tool is augmented with CloudKick for better cloud monitoring.

Table 3 Available performance metrics

Sl. No.	Performance metric	Description
Scalability based metrics		
1	CPU capacity	CPU's speed in flops
2	Memory size	In general, cache memory size for a VM
3	Scale up	Maximum number of VMs allocated for an user
4	Scale down	Minimum number of VMs allocated for an user
5	Boot time	Booting time for a VM to get ready for usage
6	Storage capacity	Storage size of data
7	Scale uptime	Time taken for increasing a specific number of VMs
8	Scale downtime	Time taken for decreasing a specific number of VMs
9	Autoscale	Boolean value for autoscaling feature
10	Response time	Time required to complete and receive a process
Architecture specific metrics		
11	Pipeline stalls	Processor specific pipeline stalls e.g. IA64
12	Cache misses	L2 or L3 cache misses
13	Frequent voltage switches	Voltage variations caused due to applications in processors such as Nehalem
Programming model based metrics		
14	MPI communication latency	Specific to MPI programs
15	Late sender or late receiver	Specific to MPI programs
16	Sequential overhead	Specific to threaded programs such as OpenMP or OpenCL

- Periscope—Periscope is a powerful distributed performance analysis tool using agents for HPC applications. The distributed nature of agents of Periscope for analyzing performance problems and the succinct information about the performance problems which are highlighted by the tool could be utilized by other existing cloud monitoring tools.
- pTop—pTop is a tool that monitors energy consumption of applications when executed in a system. The energy consumption will be a serious threat for distributed systems or distributed technologies. pTop could be integrated with other cloud monitoring tools for monitoring the energy consumption of applications as well.

4.3 Performance analysis metrics

To the best of the domain knowledge obtained due to a wide literature survey on cloud-based performance analysis methodologies and tools, the performance analysis metrics useful for analyzing the cloud resources or HPC Cloud applications are listed as given in Table 3 along with their descriptions.

Although there are many other performance metrics under the divisions, namely, architectures and programming models [30,8], the list in Table 3 could inspire cloud developers to slate their needs.

4.4 Performance monitoring of hardware counters: a challenge

Cloud-based PA tools, as discussed in Sect. 4, provide options to users to select performance metrics—typically the performance metrics from the list as shown in Table 3. In general, scalability-based metrics and programming model based metrics which identifies number of VMs or CPU speed are not so difficult to monitor because they are not much affected by virtualization. However, performance monitoring of the hardware-specific metrics is a challenge, especially when the hardware counters of virtual machines that are invisible one another were monitored. This challenge is an arousing dissent among PA tool developers in recent years.

Most of the traditional PA tools that are supported for the HPC community relies much on hardware counters for measuring performance data in a per thread basis. These hardware counters are responsible for providing information about the hardware events which are specific to the architectures. On multiple virtual machines, when two or more operating systems were booted, the performance monitoring of hardware events via hypervisors needs to be modified to get accurate measurements. Otherwise, each operating system owned by the respective VMs (guest OSs) could directly receive hardware events that need not be relevant to their VM.

Ruslan et al. [39] has explained an approach how to modify xen hypervisors to provide access to hardware performance counters in virtualized environments.

5 Conclusion

Cloud computing has proven a fertile area of work for researchers in various domains including HPC application development in cloud. Although HPC application developers have widened their awareness of utilizing clouds, they face challenges due to the resource outages, the SLA violations caused by service providers, the memory management issues, energy issues, scalability issues, and scarcity of efficient performance analysis tools.

This paper surveyed existing HPC-based cloud applications. Exploring the performance issues of solving HPC applications on cloud, the paper discussed the available performance analysis tools that pinpoint the underlying performance bottlenecks of HPC cloud applications. In the future, we intend to widely investigate on the energy monitoring issues of cloud resources.

Acknowledgments This research work is supported in part by the financial support provided by the Returning Experts programme of CIMOnline. The author appreciates the many discussions with and critical insights provided by Prof. Dr. Michael Gerndt of Technische Universitat Muenchen during his PostDoc tenure in TUM, Germany. In addition, the author thanks Shri. S. Sudershan Rao, Scientist of the Department of Science and Technology, India, and the reviewers of this survey paper for nourishing this work.

References

1. Alexandru I, Simon O, Nezih YM, Radu P, Thomas F, Epema Dick HJ (2011) Performance analysis of cloud computing services for many-tasks scientific computing. IEEE Trans Parallel Distributed Comput 22(6):931–944

2. Amazon's Cloud (2011). http://www.itnews.com.au/News/153451,stress-tests-rain-on-amazons-cloud.aspx. Accessed 31 Aug 2012
3. Amazon Cloud Watch. http://aws.amazon.com/cloudwatch/. Accessed 31 Aug 2012
4. Ang L, Xiaowei Y, Ming Z (2011) Comparing public-cloud providers. IEEE Internet Comput 15(2): 50–53
5. Anton B, Rajkumar B, Young CL, Albert Z (2012) A taxonomy and survey of energy-efficient data centers and cloud computing systems. Technical, Report, CLOUDS-TR-2010-3. arXiv:1007.0066v2
6. Arto O, Pasi T (2011) Developing a cloud business models: a case study on cloud gaming. IEEE Softw/IEEE Comput Soc 28(4):42–47
7. Ashino Y, Nakae M (2012) Virtual machine migration method between different hypervisor implementations and its evaluation. In: Proceedings of International Conference on AINA, pp 1089–1094
8. Balaji P, Buntinas D, Goodell D, Gropp W, Hoefler T, Kumar S, Lusk E, Thakur R, Traeff JL (2011) MPI on millions of cores. Parallel Process Lett (PPL) 21(1):45–60
9. Birkenheuer G, Brinkmann A, Kaiser J, Keller A, Keller M, Kleineweber C, Konersmann C, Niehoïster O, Schaïfer T, Simon J, Wilhelm M (2012) Virtualized HPC: a contradiction in terms? Softw Pract Exper 42: 485–500
10. Borenstein N, Blake J (2011) Cloud computing standards: where's the beef? Internet Comput IEEE 15(3):74–78
11. Bungo J (2011) Embedded systems programming in the cloud: a novel approach for academia. IEEE Potentials 30(1):17–23
12. Lewis N (2011) AT and T, Accenture service stores medical images in cloud. http://www.informationweek.com/healthcare/interoperability/att-accenture-service-stores-medical-ima/232200581. Accessed 31 Aug 2012
13. CloudKick Tool. https://www.cloudkick.com/. Accessed 31 Aug 2012
14. CloudStatus Tool. http://www.download.hyperic.com/pdf/cloudstatus.pdf. Accessed 31 Aug 2012
15. Díaz-Sánchez Dl, Almenarez F, Marín A, Proserpio D, Cabarcos PA (2011) Media cloud: an open cloud computing middleware for content management. IEEE Trans Consumer Electron 57(2):970–978
16. De Chaves SA, Uriarte RB, Westphall CB (2011) Toward an architecture for monitoring private clouds. IEEE Commun Magazine 49(12):130–137
17. Ganglia Tool. http://www.ganglia.info/. Accessed 31 Aug 2012
18. Grobauer B, Walloschek T, Stocker E (2011) Understanding cloud computing vulnerabilities. Secur Privacy IEEE 9(2):50–57
19. Gupta A, Milojicic D (2012) Evaluation of HPC applications on cloud. Technical Reports, HP laboratories. http://www.hpl.hp.com/techreports/2011/HPL-2011-132.html. Accessed 31 Aug 2012
20. Hong-Ling T, Schahram D (2010) Cloud computing for small research groups in computational science and engineering: current status and outlook. Computing 91(1):75–91. doi:10.1007/s00607-010-0120-1
21. Hossfeld T, Schatz R, Varela M, Timmerer C (2012) Challenges of QoE management for cloud applications. IEEE Commun Magazine 50:28–36
22. Ian F, Yong Z, Ioan R, Shiyong L (2008) Cloud computing and grid computing 360-degree compared. In: Proceedings of Grid Computing Environments Workshop, GCE '08, doi:10.1109/GCE.2008.4738445, pp 1–10
23. Information Manager Tool. http://opennebula.org/documentation:archives:rel2.0:img. Accessed 31 Aug 2012
24. InterMapper Tool. http://www.intermapper.com/about-us/news-details.aspx?newsid=26. Accessed 31 Aug 2012
25. Brandic I, Dustdar S (2011) Grid vs cloud: a technology comparison. Informat Technol 53(4):173–179. doi:10.1524/itit.2011.0640
26. Vöckler J-S, Juve G, Deelman E, Rynge M, Berriman B (2011) Experiences using cloud computing for a scientific workflow application. In: Proceedings of the 2nd international workshop on Scientific cloud, computing, ScienceCloud11. doi:10.1145/1996109.1996114
27. Juve G, Deelman E (2010) Scientific workflows and clouds. ACM Crossroads 16(1):14–18
28. Kumar K, Yung-Hsiang Lu (2010) Cloud computing for mobile users: can offloading computation save energy? Computer 43(4):51–56
29. Ye K, Che J, He Q, Huang D, Jiang X (2012) Performance combinative evaluation from single virtual machine to multiple virtual machine systems. Int J Numer Anal Model 9(2):351–370

30. Kishor K, Donghoon K, Torsten H, Frank M (2012) Assessing HPC failure detectors for MPI jobs. Accepted at 20th Euromicro International Conference on Parallel. Distributed and Network-Based Computing, Munich, Germany
31. Knight D, Shams K, Chang G, Soderstrom T (2012) Evaluating the Efficacy of the Cloud for Cluster Computation. In: Proceedings of IEEE Aerospace Conference, pp 1–10
32. LogicMonitor Tool. http://www.logicmonitor.com/quick-tour/hosted-monitoring-architecture/. Accessed 31 Aug 2012
33. Luis M, Vaquero Luis Rodero-Merino, Morán Daniel (2010) Locking the sky: a survey on IaaS cloud security. Computing 91(1):93–118. doi:10.1007/s00607-010-0140-x
34. Bull M, Hill J, Simpson A (2009) A survey of HPC systems and applications in Europe. www.prace-project.eu/IMG/pdf/Bull_DEISAPRACE.pdf. Accessed 31 Aug 2012
35. Mell P, Grance T. The NIST definition of cloud computing. http://www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf. Accessed 31 Aug 2012
36. Microsoft Cloud Outages. http://www.rscsolutions.com/news/post/microsoft-explains-recent-cloud-outage. Accessed 31 Aug 2012
37. Nagios Tool. http://www.nagios.org/. Accessed 31 Aug 2012
38. Narasimhan B, Nichols R (2011) State of cloud applications and platforms: the cloud adopters' view. Computer 44(3):24–28
39. Nikolaev R, Back G (2011) Perfctr-Xen: a framework for performance counter virtualization. Proc 7th ACM SIGPLAN/SIGOPS Int Conf Virtual Execut Environ 46(7):15–26
40. Nimsoft Tool. http://www.nimsoft.com/solutions/nimsoft-monitor/cloud.html. Accessed 31 Aug 2012
41. Oliker L, Canning A, Carter J, Shalf J, Ethier S (2004) Scientific computations on modern parallel vector systems. In: Proceedings of SC04 International Conference for High Performance Computing, Networking, Storage, and Analysis, Pittsburgh, pp 6–12
42. Ranabahu A, Anderson P, Sheth A (2011) The cloud agnostic e-science analysis platform. IEEE Internet Comput 15(6):85–89
43. Rehr JJ, Vila FD, Gardner JP, Svec L, Prange M (2010) Scientific computing in the cloud. Comput Sci Eng 12(3):34–43
44. Schaffer HE (2009) X as a service, cloud computing, and the need for good judgment. IT Professional 11(5):4–5
45. Shajulin B. HPCCLoud Research Laboratory for Tool Development. http://www.sxcce.edu.in/hpccloud. Accessed 31 Aug 2012
46. Sakr Sherif, Liu Anna, Batista Daniel M, Alomari Mohammad (2011) A survey of large scale data management approaches in cloud environments. IEEE Commun Surv Tutor 13(3):311–335
47. Ortiz S Jr (2011) The problem with cloud-computing standardization. IEEE Computer 44(7):13–16
48. Sobie RJ, Agarwal A, Anderson M, Armstrong P, Fransham K, Gable I, Harris D, Leavett-Brown C, Paterson M, Penfold-Brown D, Vliet M, Charbonneau A, Impey R, Podaima W (2011) Data intensive high energy physics analysis in a distributed cloud. http://arxiv.org/abs/1101.0357. Accessed 31 Aug 2012
49. Verena K, Debabrata D, Gre gory F, Sofia K, Anastasia A (2011) Optimal service pricing for a cloud cache. IEEE Trans Knowledge Data Eng 23(9):1345–1358
50. Windows Azure Tool. http://archive.msdn.microsoft.com/wazdmon. Accessed 31 Aug 2012