International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 5, May 2014)

CACHING as a Service: A Novel Elastic Cache Approach for Cloud Computing

Praveen M Hiremath[1]
Computer Science and Engineering
KBN College of Engineering
Gulbarga, Karnataka
praveenmhh000@gmail.com

Afroze Ansari[2]
Asst Professor, Computer Science and Engineering
KBN College of Engineering
Gulbarga, Karnataka
ansariafroze@yahoo.com

***Abstract*— Cloud computing plays a vital role in today's distributed systems widely used by internet users. It provides a flexible and consistent environment in which the data, devices and services can be shared among end users in order to save the time and cost. The advanced computing is the most important service provided by this technology. Due to the complexity and difficulty of services requested by IT industries and organizations, cloud computing has become a widely adopted technology to expedite and facilitate the process of service delivery through internet. In addition to technical challenges, providing cache services in clouds encounters a major practical issue (quality of service or service level agreement issue) of pricing. Currently, (public) cloud users are limited to a small set of uniform and coarse-grained service offerings, such as High-Memory and High-CPU in Amazon, EC2. In this paper, we present the cache as a service (CaaS) model as an optional service to typical infrastructure service offerings. Specifically, the cloud provider sets aside a large pool of memory that can be dynamically partitioned and allocated to standard infrastructure services as disk cache. A completely redesigned D-cache variant that is more effective (provides tighter lower/upper bounds) and also more efficient (faster bound determination) than the previous version is proposed.**

***Index Terms*— Cloud Computing, Cache, CaaS, Remote memory, Cost efficiency.**

I. INTRODUCTION

Those who have an interest in information technology may often come across the term 'cloud computing'. Despite its vagueness and the fact that it is too broad to be discussed, there is an agreement among IT professionals that 'cloud computing' is a technical concept, where system users save the information on remote servers which are managed by others, and use the applications that are stored inside the server and executed from other locations, instead of from their own computers. The internet plays an important role for the development of a variety of technologies. There is no doubt that one of the most commonly discussed topics related to technologies is cloud computing.

In computer networking, cloud computing is a phrase used to describe a variety of computing concepts that involve a large number of computers connected through a real-time communication network such as the internet. It is very similar to the concept of utility computing. In science, cloud computing is a synonym for distributed computing over a network, and means the ability to run a program or application on many connected computers at the same time.

The phrase is often used in reference to network-based services, which appear to be provided by real server hardware, and are in fact served up by virtual hardware, simulated by software running on one or more real machines. Such virtual servers do not physically exist and can therefore be moved around and scaled up or down on the fly without affecting the end user, Somewhat like a cloud becoming larger or smaller without being a physical object.

In common usage, the term "the cloud" is essentially a metaphor for the internet. Marketers have further popularized the phrase "in the cloud" to refer to software, platforms and infrastructure that are sold "as a service", i.e. remotely through the internet. Typically, the seller has actual energy-consuming servers which host products and services from a remote location, so end-users don't have to; they can simply log on to the network without installing anything. The major models of cloud computing service are known as software as a service, platform as a service, and infrastructure as a service. These cloud services may be offered in a public, private or hybrid network. Google, Amazon, oracle cloud, Salesforce, Zoho and Microsoft azure are some well-known cloud vendors.

A systematic literature review is often conducted to explore the key aspects of a scientific concept. It can be helpful to be employed by researchers as a research roadmap or an initial reference. Indeed, it expedites the process of problem finding that is a critical part of any research [1-10].

II. PROPOSED METHOD

Cloud computing relies on sharing of resources to achieve coherence and economies of scale, similar to a utility (like the electricity grid) over a network. At the foundation of cloud computing is the broader concept of converged infrastructure and shared services. The cloud also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shared by multiple users but are also dynamically reallocated per demand. This can work for allocating resources to users. For example, a cloud computer facility that serves European users during European business hours with a specific application (e.g., email) may reallocate the same resources to serve north American users during north America's business hours with a different application (e.g., a

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 5, May 2014)

web server). This approach should maximize the use of computing powers thus reducing environmental damage as well since less power, air conditioning, rack space, etc. With cloud computing, multiple users can access a single server to retrieve and update their data without purchasing licenses for different applications. The term "moving to cloud" also refers to an organization moving away from a traditional CAPEX model (buy the dedicated hardware and depreciate it over a period of time) to the OPEX model (use a shared cloud infrastructure and pay as one uses it).[11].

Proponents claim that cloud computing allows companies to avoid upfront infrastructure costs, and focus on projects that differentiate their businesses instead of infrastructure. Proponents also claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less maintenance, and enables it to more rapidly adjust resources to meet fluctuating and unpredictable business demand. Cloud providers typically use a "pay as you go model." this can lead to unexpectedly high charges if administrators do not adapt to the cloud pricing model. In this paper, we present the cache as a service (CaaS) model as an optional service to typical infrastructure service offerings. Specifically, the cloud provider sets aside a large pool of memory that can be dynamically partitioned and allocated to standard infrastructure services as disk cache.

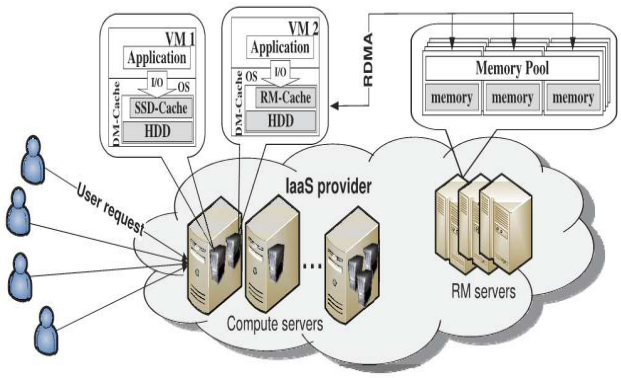

Fig: A block diagram of CAAS model.

The CaaS model consists of two main components: an elastic cache system as the architectural foundation and a service model with a pricing scheme as the economic foundation. The basic system architecture for the elastic cache aims to use RM, which is exported from dedicated memory servers (or possibly ssds). It is not a new caching algorithm. The elastic cache system can use any of the existing cache replacement algorithms. Near uniform access time to RM-based cache is guaranteed by a modern high speed network interface that supports RDMA as primitive operations. Each VM in the cloud accesses the RM servers via the access interface that is implemented and recognized as a normal block device driver. Based on this access layer, VMS utilize RM to provision a necessary amount of cache memory on demand. A group of dedicated memory servers exports their local memory to VMs, and exported memory space can be viewed as an available memory pool as shown in Fig. 1.

This memory pool is used as an elastic cache for VMs in the cloud. For billing purposes, cloud service providers could employ a lease mechanism to manage the RM pool. To employ the elastic cache system for the cloud, service components are essential. The CaaS model consists of two cache service types (CaaS types) based on whether LM or RM is allocated with. Since these types are different in their performance and costs a pricing scheme that incorporates these characteristics is devised as part of CaaS.

An approach for general metric access methods based on D-cache (distance cache), that helps to reduce the cost of both, indexing and querying is being applied on elastic system. The basic task of D-cache is to determine cheaper tight lower- and upper bound of an unknown distance between two objects, based on stored distances computed during previous querying and/or indexing. Although the D-cache was already introduced in our preliminary work, it was applied in a more narrowed context—as a tool for efficient index-free similarity search (resulting in a new method, the D-file). Moreover, in this paper we not only employ the D-cache in various MAMs, but we present a completely redesigned D-cache variant that is more effective (provides tighter lower/upper bounds) and also more efficient (faster bound determination) than the previous version.

III. EXPERIMENTAL RESULTS

Among many important factors in designing an elastic cache system, we particularly focus on the type of cache medium, the implementation level of our cache system, the communication medium between a cache server and a VM, and reliability. Elastic cache can be deployed at either application or OS level (block device or file system level). In this paper, it is the fundamental principle that the cache need not affect application code or file systems owing to the diversity of applications or file system configurations on cloud computing. Application level elastic cache such as mem-cached could have better performance than OS level cache, since application level cache can exploit application semantics. However, modification of application code is always necessary for application level cache [12].

A file system level implementation can also provide many chances for performance improvements, such as buffering and pre-fetching. However, it forces users to use a specific file system with the RM-based cache. In contrast, although a block-device level implementation has fewer chances of performance improvements than the application or file system level counterpart, it does not depend on applications or file systems to take benefits from the underlying block-level cache implementation. One of most important requirements for the elastic cache is failure resilience. Since we implement the elastic cache at the block device level, the cache system is designed to support a RAID-style fault-tolerant mechanism[13].Based on a RAID-like policy, the elastic cache can detect any failure of cache servers and recovers automatically from the failure (a single cache server failure).

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 5, May 2014)

The elastic cache system is conceptually composed of two components: a VM and a cache server. A VM demands RM for use as a disk cache. We build an RM-based cache as a block device and implement a new block device driver (RM-Cache device). In the RM-Cache device, RM regions are viewed as byte-addressable space. The block address of each block I/O request is translated into an offset of each region, and all read/write requests are also transformed into RDMA read/write operations. We use the device-mapper module of the Linux operating system (i.e., DM-Cache) to integrate both the RM-Cache device and a general block device (HDD) into a single block device [14-15]. This forms a new virtual block device, which makes our cache pluggable and file-system independent. The Fig. 2 shows elastic cache structure and double paging problem.

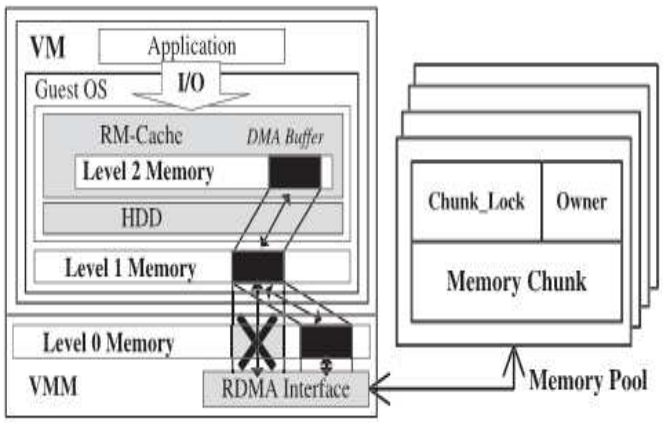

Fig2: Elastic cache structure and double paging system for the proposed CaaS model.

In our CaaS model, it is assumed that a user, who sends a request with a CaaS option (HP or BV), also accompanies an application profile including data volume, data access pattern, and data access type. D-cache (distance cache) based model proposed in this paper helps to reduce the cost of both, indexing and querying. It can be argued that these pieces of application specific information might not be readily available particularly for average users, and some applications behave unpredictably. In this paper, we primarily target the scenario in which users repeatedly and/or regularly run their applications in clouds, and they are aware of their application characteristics either by analyzing business logic of their applications or by obtaining such information using system tools. A pricing model that explicitly takes into account various elastic cache options is essential for effectively capturing the tradeoff between (I/O) performance and (operational) cost. A tool for efficient index-free similarity search is defined.

For performance evaluation, we used a7-node cluster, each node of which is equipped with an Intel(r) Core(tm)2 Quad CPU 2.83 GHz and 8 GB Ram. All nodes are connected via both a switched 1 Gbps Ethernet and 10 Gbps Infiniband. A memory server runs Ubuntu 8.0.4 with Linux 2.6.24 kernel and exports 1 GB memory.

IV. CONCLUSION

The cost efficiency of CaaS is evaluated through extensive simulations with randomly generated workloads, and each simulation is conducted using the metric for performance improvement of each cache. Different workload characteristics were applied. With the increasing popularity of infrastructure services such as Amazon EC2 and Amazon RDS, low disk I/O performance is one of the most significant problems. In this paper, we have presented a CaaS model as a cost efficient cache solution to mitigate the disk I/O problem in IaaS. To this end, we have built a prototype elastic cache system using a remote-memory-based cache, which is pluggable and file system independent to support various configurations. This elastic cache system together with the pricing model devised in this study has validated the feasibility and practicality for CaaS model. Through extensive experiments, we have confirmed that CaaS helps IaaS improve disk I/O performance greatly The performance improvement gained using cache services clearly leads to reducing the number of(active) physical machines the provider uses, increases throughput, and in turn results in profit increase. This profitability improvement enables the provider to adjust its pricing to attract more users. Use of D-cache has reduced cost incurred on indexing and querying in this model.

REFERENCES

[1]. Christos Kalloniatis, Haralambos Mouratidis, Manousakis Vassilis, Shareeful Islam, Stefanos Gritzalis, Evangelia Kavakli, towards the design of secure and privacy-oriented information systems in the cloud: Identifying the major concepts, Computer Standards & Interfaces, Volume 36, Issue 4, June 2014, Pages 759-775, ISSN 0920-5489.

[2]. Ahmed Patel, Mona Taghavi, Kaveh Bakhtiyari, Joaquim Celestino Júnior, An intrusion detection and prevention system in cloud computing: A systematic review, Journal of Network and Computer Applications, Volume 36, Issue 1, January 2013, Pages 25-41, ISSN 1084-8045.

[3]. Giuseppe Aceto, Alessio Botta, Walter de Donato, Antonio Pescapè, Cloud monitoring: A survey, Computer Networks, Volume 57, Issue 9, 19 June 2013, Pages 2093-2115, ISSN 1389-1286.

[4]. Amin Jula, Elankovan Sundararajan, Zalinda Othman, Cloud computing service composition: A systematic literature review, Expert Systems with Applications, Volume 41, Issue 8, 15 June 2014, Pages 3809-3824, ISSN 0957-4174.

[5]. Mark D. Ryan, Cloud computing security: The scientific challenge, and a survey of solutions, Journal of Systems and Software, Volume 86, Issue 9, September 2013, Pages 2263-2268.

[6]. Georgia Sakellari, George Loukas, A survey of mathematical models, simulation approaches and testbeds used for research in cloud computing, Simulation Modelling Practice and Theory, Volume 39, December 2013, Pages 92-103.

[7]. Vanessa Ratten, Cloud computing: A social cognitive perspective of ethics, entrepreneurship, technology marketing, computer self-efficacy and outcome expectancy on behavioural intentions, Australasian Marketing Journal (AMJ), Volume 21, Issue 3, August 2013, Pages 137-146, ISSN 1441-3582.

[8]. Angela Lin, Nan-Chou Chen, Cloud computing as an innovation: Percepetion, attitude, and adoption, International Journal of Information Management, Volume 32, Issue 6, December 2012, Pages 533-540, ISSN 0268-4012.

[9]. Dawei Sun, Guiran Chang, Lina Sun, Xingwei Wang, Surveying and Analyzing Security, Privacy and Trust Issues in Cloud

Computing Environments, Procedia Engineering, Volume 15, 2011, Pages 2852-2856, ISSN 1877-7058.

[10]. Luís Monteiro, André Vasconcelos, Survey on Important Cloud Service Provider Attributes Using the SMI Framework, Procedia Technology, Volume 9, 2013, Pages 253-259, ISSN 2212-0173.

[11]. Nabil Sultan, Cloud computing: A democratizing force?, International Journal of Information Management, Volume 33, Issue 5, October 2013, Pages 810-815, ISSN 0268-4012.

[12]. Hyuck Han; Young Choon Lee; Woong Shin; Hyungsoo Jung; Yeom, H.Y.; Zomaya, A.Y., "Cashing in on the Cache in the Cloud," Parallel and Distributed Systems, IEEE Transactions on , vol.23, no.8, pp.1387,1399, Aug. 2012

[13]. Rimal, B.P.; Eunmi Choi; Lumb, I., "A Taxonomy and Survey of Cloud Computing Systems," INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on , vol., no., pp.44,51, 25-27 Aug. 2009, doi: 10.1109/NCM.2009.218.

[14]. Minqi Zhou; Rong Zhang; Wei Xie; Weining Qian; Aoying Zhou, "Security and Privacy in Cloud Computing: A Survey," Semantics Knowledge and Grid (SKG), 2010 Sixth International Conference on , vol., no., pp.105,112, 1-3 Nov. 2010.

[15]. Patidar, S.; Rane, D.; Jain, P., "A Survey Paper on Cloud Computing," Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on , vol., no., pp.394,398, 7-8Jan. 2012, doi: 10.1109/ACCT.2012.15.