Building Crawler Engine on Cloud Computing Infrastructure

Sinung Suakanto[#1], Suhono H. Supangkat[#2], Suhardi[#3], Roberd Saragih[#4], I Gusti Bagus Baskara Nugraha[#5]
[#1,2,3,5]School of Electrical Engineering and Informatics , Institut Teknologi Bandung
Jalan Ganesha 10 Bandung 40132, Indonesia
[#4]Mathematic Department, Institut Teknologi Bandung
Jalan Ganesha 10 Bandung 40132, Indonesia
Email: [1]sinung@students.itb.ac.id, [2]suhono@stei.itb.ac.id, [3]suhardi@stei.itb.ac.id, [4]roberd@math.itb.ac.id, [5]baskara@stei.itb.ac.id

Abstract-This paper is aimed to create implementation crawler engine or search engine using cloud computing infrastructure. This approach use virtual machines on a cloud computing infrastructure to run service engine crawlers and also for application servers. Based on our initial experiments, this research has successfully built crawler engine that runs on Virtual Machine (VM) of cloud computing infrastructure. The use of Virtual Machine (VM) on this architecture will help to ease setup or installation, maintenance or VM terminating that has been running with some particular service crawler engine as needed. With this infrastructure, the increasing or decreasing in capacity and capability of multiple engine crawlers could set easily and more efficiently.

Keyword: crawler engine, search engine, cloud computing, efficient infrastructure

I. INTRODUCTION

Nowadays search engines have influenced people's behavior in using the Internet. Various search engines have been built up and some of them are very dominant, like Google, Yahoo, and Bing. Google, for example, logs 2 billion searches per day and 300 million users use the search facility provided by Google on a daily basis [1]. This number is set to rise in the future so that the facility must be provided by very huge infrastructure and must be coordinated by each other. The infrastructure in this context could consist of a number of machines that runs crawler engine, application servers to accommodate the high demand and storage servers or disk to store the results of all searches.

One of the main problems of search engine service is the requirement of large computing resources to produce a complete and quick search. In practice, the service search engine require big of infrastructures like a lot of servers and big storage to run its services. The issue of search engine infrastructure management becomes important when the number of infrastructure continues to increase. The use of effective and efficient infrastructure is important enough so there isn't resources that having over or under utilities. Several studies have developed faster and more efficient search engines [1], [2], [3].

Cloud computing claimed that it provides better efficiency in the use of infrastructure [6]. In addition, cloud computing technology has been developed so rapidly and could change the implementation or operation mode in Information Communication Technology. By using cloud computing, the use of information technology infrastructure is claimed to be more efficient and more effective. Therefore, this study aims to establish a system of search engines by using cloud computing infrastructure.

II. RELATED WORK

In this section we will describe the state of the art of research on search engine. Some of them describe about crawler engine itself and how it work. Several studies have been developed to developing search engine method to contribute to scientific contribution or industrial-scale use.

Some of specific crawler developed for special purposes. Topical or focused web crawlers were developed to create contextual search engine or more focused result [2], [3]. These crawlers automatically navigate the hyperlinked structure of the web while using link contexts to predict the benefit of following the corresponding hyperlinks with respect to some initiating topic or theme. Context of a hyperlink or link context is defined as the terms that appear in the text around a hyperlink within a web page. Link contexts have been applied to a variety of web information retrieval and categorization tasks. Topical or focused web crawlers have a special reliance on link contexts.

The basic elements of a search engine process are crawling, storage, indexing and ranking algorithms [1]. Many studies has proposes new approach provides a cloud-based platform for low-cost, effective and personalized search models [1]. Cloud computing today has brought a new era in the use of infrastructure more efficient. Currently, cloud computing infrastructure will maximize the use of virtual machine (VM). It means that we can create instance of VM(s) on a single physical computer. The use of VM is expected to increase the efficiency of resource usage because VM can be set, invoked or terminated in accordance with the requirements.

Typically, cloud computing architecture consists of the front end and the node as shown in figure 1 [6]. Front end on a specific implementation in Eucalyptus often referred or named with Cloud Controller [4],[5]. While node or Node Controller is a physical device that can run single or multiple Virtual

Machines (VMs) based on demand. The management of the number allocates for VM(s), and VM capacities perform by the front end. With this capability, then the cloud computing capacity can be managed and allocated efficiently by allocating Virtual Machine (VM) in accordance with necessary needs. For example, at the peak time session, we can determine the VM with a large capacity as the host server or we can also duplicate VM with same specifications.

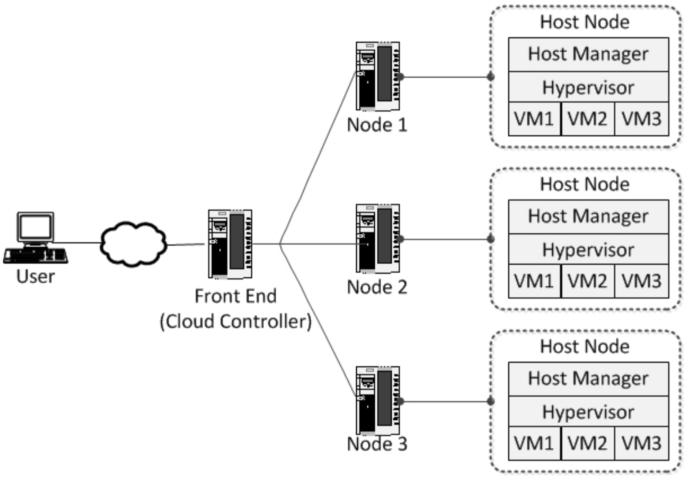


Figure 1 – Typical Cloud Computing Architecture

Key advantages of cloud computing is the use of virtualization so that the users do not need to know where the computation performed by a machine [6], [9]. Also, with the usage of VM(s) will make it easier when running application and operating system installation. By using the VM, we can easily create a master application or service built in an operating system resulting in image. If we need some similar system to run the same program then cloud computing will easily turn on or duplicate the same VM on a particular physical node computer without hard installation.

At a previous work, we had been proposed a method for maintaining quality of service for applications using cloud computing services. One of parameter being maintained is that the average response time value must be above a certain value [8]. The method was developed using the FTR-HTTP (Finite Time Response-HTTP) which is a method to determine the response time is expected to be above a certain value [8]. The concept of FTR-HTTP is proposing a simple method to guarantee service quality on the Internet TCP / IP without having to change devices or existing network configuration. The technique proposed in this study is to add a blocking mechanism in the form of restrictions on the HTTP protocol [8]. This method would be applied into search engine service in order to guarantee average response time received by users.

III. PROPOSED ARCHITECTURE

This research tries to build a search engine service in the cloud computing architecture. Even thought, it is not pure new crawler engine technique, but this study will describe how the implementation of it in cloud computing. And also, this research has two contributions:

- How to build crawler engine that running on cloud computing infrastructure.
- How to make search engine service that using guarantee mechanism so that the users could have better performance for searching.

3.1. Search Engine Works

We can define how typically search engine works as depicted on figure 2 [6]. Search Engines has main function in crawler module and indexing module. Crawler module would search on the web and save the result into page repository. Indexing Module would continue to get information from Meta, content or other intelligent methods. The indexes results would be displayed whenever any request or query from user to find specific search.

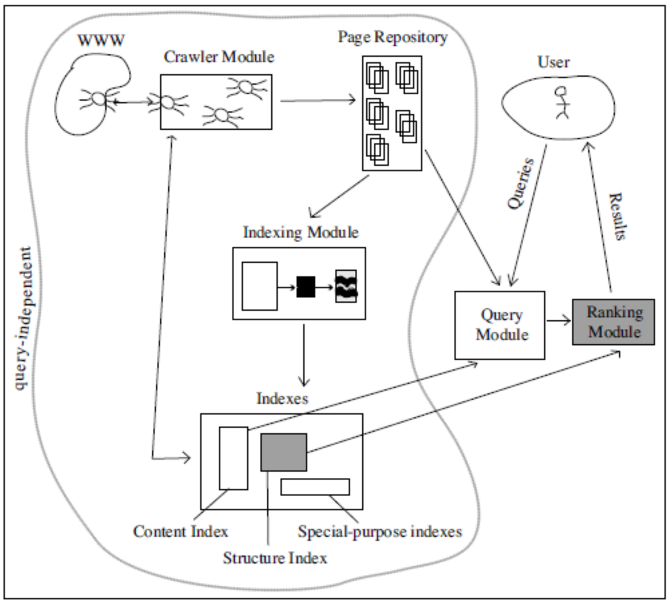


Figure 2 – Element of Search Engine

3.2 Search Engine Running on Virtual Machines

This study want to develop our own search engine that running on virtual machines. To propose this concept, we developed search engine architecture into a cloud infrastructure as shown in figure 3. Because of the basic characteristics of the infrastructure is the use of virtual machine as a worker, then we will also take advantage of the usage virtual machine optimally.

Our current architecture proposed to use Virtual Machine (VM) on cloud computing to run both of crawler engine services and application servers. Also, we build specific cloud storage to save, record and indexing the results from crawler engine. We define and design our framework to save keyword and for more effective searching.

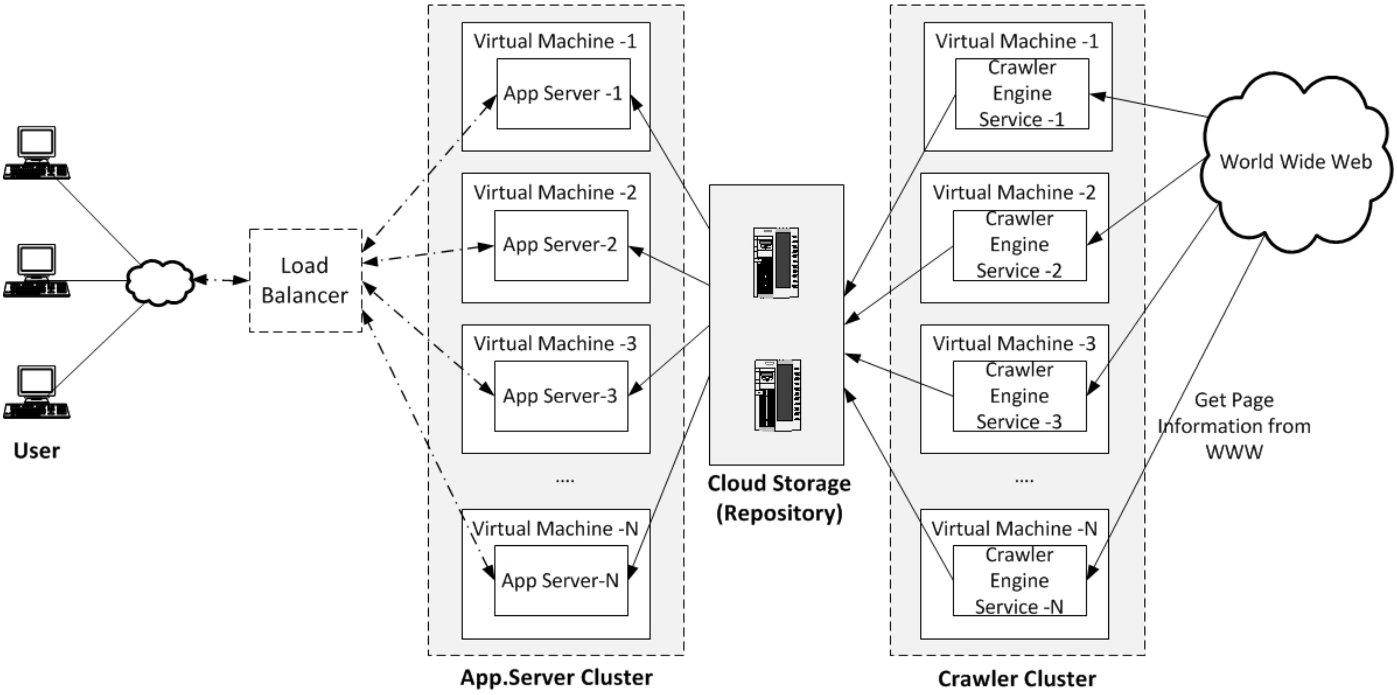


Figure 3 – Architecture of Search Engine on Cloud Computing Infrastructure

3.3. Crawler Engine Service

For crawler engine service, we build service with architecture shown on figure 4.

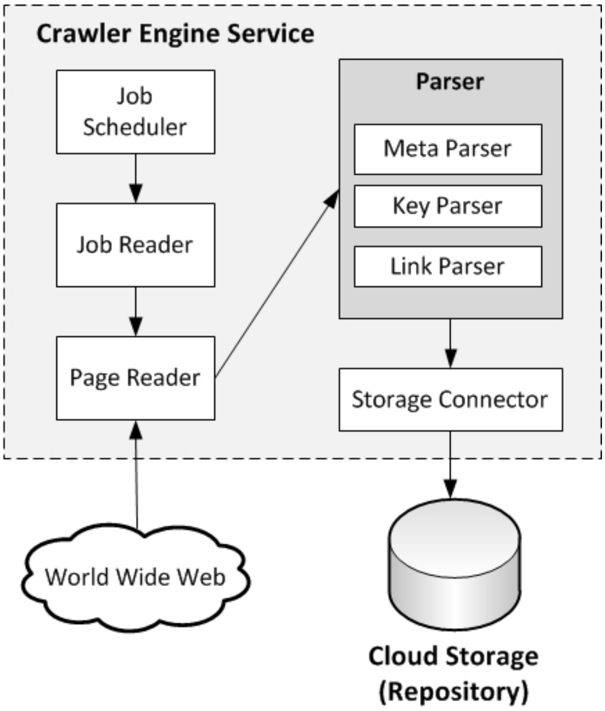


Figure 4 – Crawler Engine Service

Form the figure we can define there are 5 basic modules:

- *Job Scheduler*
  Job Scheduler has function to schedule jobs to visiting or crawling over the World Wide Web. At this module, we can implement specific algorithm to run specific purpose. The algorithm could define how the crawling can act, even bread-first or deep-first algorithm to get more precise or wider results.
- *Job Reader*
  Job Reader has function to read job that queuing and ready to be executed. Job Reader has function to start, pause, resume or stop depending on how it would be run. Job Reader also has function to set the speed to execute single job. We can define slow, middle or fast crawling with this module. This speed setting will affect such things as: how much traffic is flowing, how other hardware capabilities in handling high transaction.
- *Page Reader*
  Page Reader has function to crawling or visiting specific URL based on selected by Job Reader. This function requires a connection to the internet to visit websites that want to go. The results of these readings will be forwarded to the module Parser.
- *Parser*
  This module has the function to decompose the result from reading in forms such as HTML code to be broken down into a few things such as keywords, link or Meta. This is the basic information that we want to know and must be store for the next searching.

- *Storage Connector*
  This module has the function to store the parser results into storage. Since there are a lot of data to be stored, this module will take time to process it.

Major delay to process single job could cause by several processes such as:

- Delay from reading or visiting websites. This delay identically with response time when we are visiting website using specific browser/
- Delay from parsing data. Because the system must read and parse a huge data, so it would consume several time.
- Delay to store the parser result. Delay is caused by a number of data like keywords, pages, links has been found. As more data is found then the longer the storage process for storing process carried out through a cloud network system.

3.3. Front End Implementation

For front end, we use another mechanism proposed from our research using FTR-HTTP [8]. Our basic method using simple blocking mechanism for any specified request that doesn't fit with specific network quality parameter. If we denote P_{t+1} as next process or request on user interaction on web-client, and 1 denote as non-blocking state, 0 for blocking state, we can define:

$$P_{t+1} = \begin{cases} 1 & if(RT_t^i < RT_{\max}^i) \\ 0 & if(RT_t^i \geq RT_{\max}^i) \end{cases} \qquad (1)$$

Where:

RT_t^i : response time for event-i at time t (current)

$RT_{\max}^i$: maximum response time for event-i

For implementation on the front-end application could be seen in Figure 5. The technique can be implemented easily using AJAX techniques (Asynchronous Java Script). AJAX moves web-based applications from a page model to a true application model, based on events and user actions rather than pages [10].

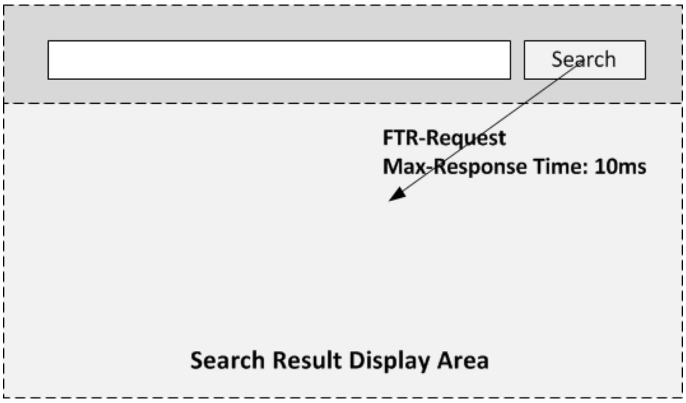
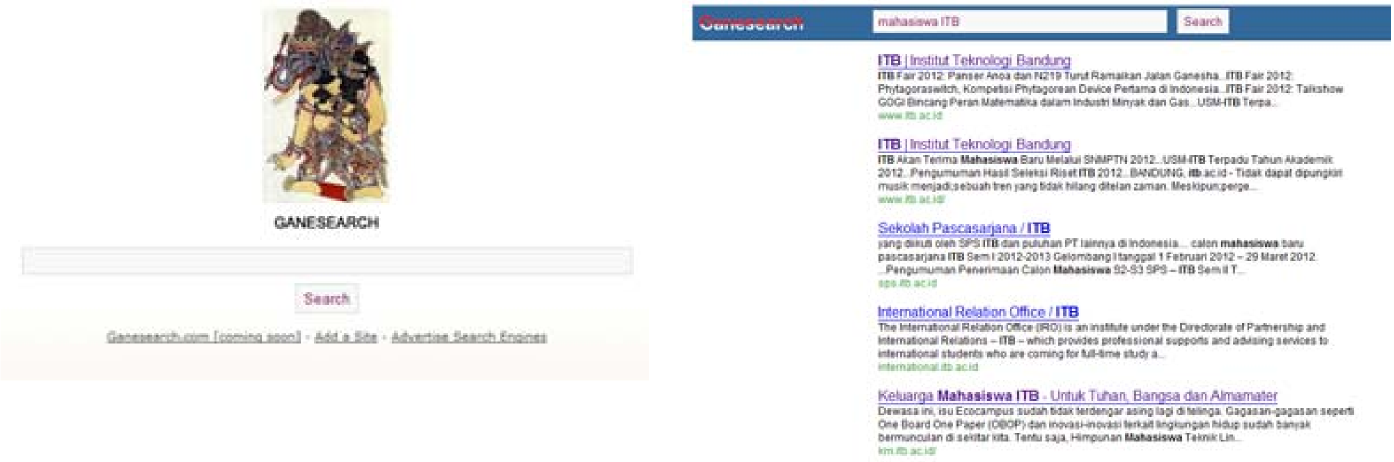


Figure 5– Front End Design using FTR-HTTP

IV. RESULT & ANALYSIS

This study has successfully implemented this search engine on cloud infrastructure and named with "*Ganesearch*". In this study, we use Eucalyptus for cloud infrastructure deployment. For the front end user interface, we can display the result for our search engines result depicted on figure 6. For our prototype implementation, we are focused on search limited only to word or string only and not yet for any resources like file, images, etc.

In the initial experiments, we observed for single crawler engine on single Virtual Machine (VM) running for 10 days of observation. Crawler engine is run at a moderate speed or even slow speed to avoid flooding of outbound traffic when accessing certain pages. Also, the actual speed is automatically adjusted based on speed for parsing data processing and delay to storage in a repository.


GANESEARCH
Search
Ganesearch.com [coming soon] · Add a Site · Advertise Search Engines
Ganesearch
mahasiswa ITB
Search

ITB | Institut Teknologi Bandung
ITB Fair 2012: Panser Anoa dan N219 Turut Ramaikan Jalan Ganesha...ITB Fair 2012: Phytagoraswitch, Kompetisi Phytagorean Device Pertama di Indonesia...ITB Fair 2012: Talkshow GOGI Bincang Peran Matematika dalam Industri Minyak dan Gas...USM-ITB Terpa...
www.itb.ac.id

ITB | Institut Teknologi Bandung
ITB Akan Terima **Mahasiswa** Baru Melalui SNMPTN 2012...USM-ITB Terpadu Tahun Akademik 2012...Pengumuman Hasil Seleksi Riset ITB 2012...BANDUNG, itb.ac.id - Tidak dapat dipungkiri musik menjadi sebuah tren yang tidak hilang ditelan zaman. Meskipun perge...
www.itb.ac.id/

Sekolah Pascasarjana / ITB
yang diikuti oleh SPS ITB dan puluhan PT lainnya di Indonesia.... calon **mahasiswa** baru pascasarjana ITB Sem I 2012-2013 Gelombang I tanggal 1 Februari 2012 – 29 Maret 2012 ...Pengumuman Penerimaan Calon **Mahasiswa** S2-S3 SPS – ITB Sem II T...
sps.itb.ac.id

International Relation Office / ITB
The International Relation Office (IRO) is an institute under the Directorate of Partnership and International Relations – ITB – which provides professional supports and advising services to international students who are coming for full-time study a...
international.itb.ac.id

Keluarga **Mahasiswa ITB** - Untuk Tuhan, Bangsa dan Almamater
Dewasa ini, isu Ecocampus sudah tidak terdengar asing lagi di telinga. Gagasan-gagasan seperti One Board One Paper (OBOP) dan inovasi-inovasi terkait lingkungan hidup sudah banyak bermunculan di sekitar kita. Tentu saja, Himpunan **Mahasiswa** Teknik Lin...
km.itb.ac.id/

(a)

(b)

Figure 6 – Front End Display Result

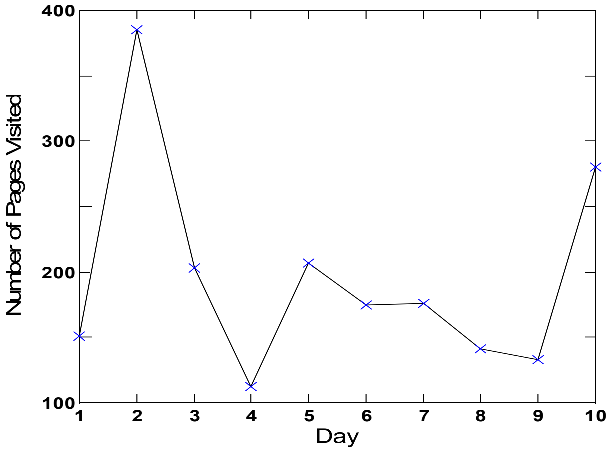


Figure 7 – Number of Pages Visited for 10 days

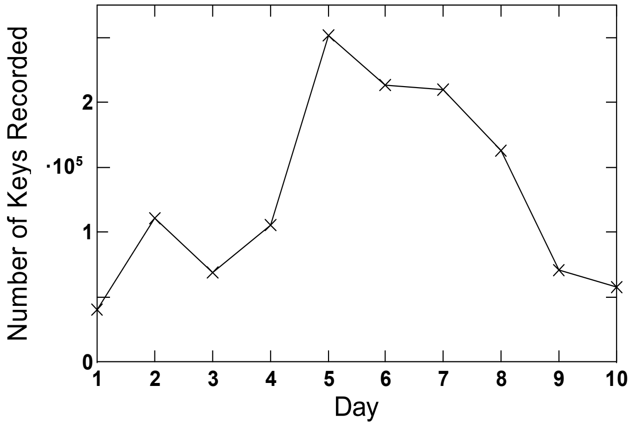


Figure 8 – Number of Keywords Recorded for 10 days

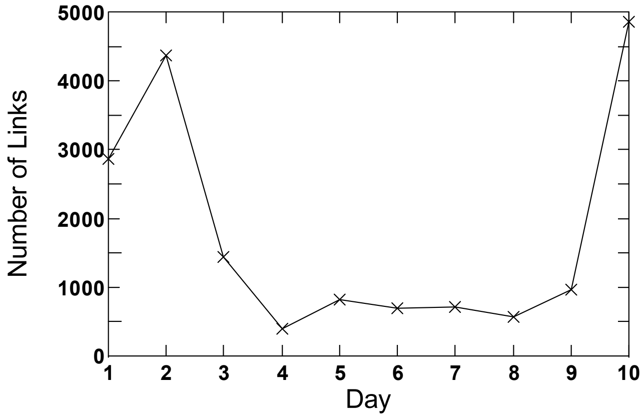


Figure 9 – Number of Links for 10 days

In this experiment, we are interested to calculate the resulting performance of a crawler on a single Virtual Machine (VM). Some of the results to be analyzed such as the number of pages are visited every day, the number of keywords that are recorded and the numbers of new links are discovered every day. These results can be seen in Figure 7-9.

Based on initial experiments and from the graph on figure7-9, for every single crawler on a single VM will generate an average visit to the other pages to be indexed by an average of 196.3 sites per day. For visit indexing will result in an average of 129,047.7 keywords and found 1,770 links each day. This speed of indexing is still quite slow compared to some existing search engines. This is because the system in experiments using setting the waiting delay between the visits made in 5 minutes.

To produce a faster search results can be used more engine crawler and mode Virtual Machine (VM) used. In cloud computing infrastructure, it is very simple to create new instance of VM that running crawler engine instead of setup a new hardware with new installation. In addition, the speed of web visiting (crawling) can be set in such a way as to produce an optimal speed. Optimal here is intended also to the servers that run the service does not have loads more.

IV. CONCLUSION & FUTURE WORK

This study has show how search engines that consist of crawler engine can be run in the cloud computing infrastructure. At this work, we have successfully developed both crawler engine and application service prototype. Further work, we will measure the performance of this search system with a higher load of searching, with more VM running. In addition, we will also measure how the impact of FTR-HTTP can be used to ensure the quality of the user.

REFERENCES

[1] Akassh A Mishra, Chinmay Kamat, "Migration of Search Engine Process into the Cloud", *International Journal of Computer Applications*, Volume 19– No.1, April 2011

[2] Gautam Pant and Padmini Srinivasan, "Link Contexts in Classifier-Guided Topical Crawlers", *IEEE Transactions On Knowledge and Data Engineering, Vol.18, No.1, January 2006.*

[3] Soumen Chakrabarti, Martin van den Berg, Byron Domc. "*Focused crawling: a new approach to topic-specific Web resource discovery*", 1999

[4] Daniel Nurmi et al, "Eucalyptus : A Technical Report on an Elastic Utility Computing Architecture Linking Your Programs to Useful Systems"

[5] White Paper , "Intel® Cloud Builder Guide to Cloud Design and Deployment on Intel® Platforms"

[6] Sheheryar Malik, Fabrice Huet, "Virtual Cloud: Rent Out the Rented Resources", 6th IEEE International Conference for Internet Technology and Secured Transactions 2011

[7] Langville A., Meyer C., Google's Page Rank and Beyond, Princeton University Press,2006

[8] Sinung Suakanto, Suhono Harso Supangkat, Suhardi, "*Introduction to Finite Time Response-HTTP: a Simplest Way to Guarantee Quality of Service of Web Application under Best Effort Network*", Proceedings International Conference AOTULE 2010

[9] Rajkumar Buyyaa, Chee Shin Yeoa, Srikumar Venugopala, James Broberg, Ivona Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility", ELSEVIER Future Generation Computer Systems, 2008

[10] Lori MacVittie , "The Impact of AJAX on the Network", White Paper

.