Wong, L.-H. et al. (Eds.) (2013). Proceedings of the 21st International Conference on Computers in Education. Indonesia: Asia-Pacific Society for Computers in Education

A Private Cloud Environment for Teaching Search Engine Construction

Eisuke ITO[a*], Brendan FLANAGAN[b], Chengjiu YIN[a], Tetsuya NAKATOH[a] & Sachio HIROKAWA[a]
[a]*Research Institute for Information Technology, Kyushu University , Japan*
[b]*Graduate School of ISEE, Kyushu University , Japan*
*ito.eisuke.523@m.kyushu-u.ac.jp

Abstract: Kyushu University installed a private cloud system, named "campus cloud system", using VCL and CloudStack. For a graduate school exercise course on web search engine, the authors prepared a virtual machine on VCL, which had apache web server and GETA indexer preinstalled. This paper introduces an outline of the cloud system, the exercise, and also reports advantages and disadvantages of cloud based education.

Keywords: Cloud computing, private cloud, virtual machine, VCL, exercise, web search engine

1. Introduction

In recent years, cloud technologies are used in computer-aided education of universities, and particularly in information science and technology education. Kyushu University installed a private cloud system in 2011, named "Kyu(Q)shu University Campus Cloud System (Qcloud for short)", using VCL [1,2,3,4] and CloudStack [5] for graduate schools education and research [6,7].

VCL (Virtual Computing Lab.) realizes DaaS (Desktop as a Service) cloud service. It was developed by NCSU (North Carolina State University) in 2004 [3,4]. NCSU has donated the VCL source code to the ASF (Apache Software Foundation), and it is now provided as open source software. In VCL, a teacher prepares a customized virtual machine (VM for short), which is installed with specific applications for a lecture. A user can request a VM through the VCL web interface, upon which VCL copies the VM image to a real machine and starts it. After starting the VM, a student can access the VM using RDP (remote desktop protocol) or SSH (secure shell), and use the applications.

We have been providing a course on web search engine for graduate students of the school of ISEE (information science and electrical engineering), Kyushu University. Most lectures in the course were discourse style until 2011. It was difficult to teach real technologies and skills by classroom lectures. Thereby, student satisfaction for the course was low. In 2012, to drastically improve the course, we decided to introduce exercises in which an actual search engine is made in the lecture using the VCL system. For use in the exercises of the course, we prepared a Linux VM on VCL with Apache web server and GETA [8] indexer preinstalled.

The present paper reports the system, the lecture, the exercises and the lessons learned. The composition of this paper is as follows. In section 2, we briefly introduce the motivation behind the Qcloud system installation, and an outline of Qcloud. Section 3 describes the web search engine exercise in which students used the Linux VM on VCL to construct a web search engine. In section 4, we describe the exercise steps, and report student feedback that was collected by questionnaire. We also discuss the advantages and disadvantages of cloud based education. Finally, we conclude this paper in section 5.

2. Kyushu University Campus Cloud System

This section briefly describes the Qcloud system.

Motivation

The graduate school of ISEE has been operating a computer system for education and research. Until 2011, the ISEE provided an on-premises computer system that was installed in the ISEE server room. ISEE headquarters had decided on a system replace plan, within which it indicated the intention to replace the existing on-premise system with a cloud-computing system. According to the plan, ISEE decided to use two cloud-computing systems, one was a public cloud service such as AWS (amazon web service), and the other was to use a private cloud operated by RIIT (research institute for information technology), Kyushu University. Over the course of half a year, we discussed the design and structure of the private cloud system, Qcloud. As a result, we decided to introduce four sub-cloud systems as shown in Table 1.

Table 1: Sub cloud systems in Qcloud.

Name	Description
Education	Lecture and exercise in graduate school. Like a PC classroom.
Sever	Commodity service servers such as mail or web. Long-term period.
Development	Software or service system development. Short-term period.
Data Processing	Big data processing.

An outline of Qcloud

Figure 1 shows an outline of the architecture of Qcloud. We started trial use of Qcloud in October 2012, and only started to provide Qcloud to ISEE as of April 2013. We are planning to provide Qcloud to other departments and graduate schools in the future. In addition, we are also planning to provide it to other institutes.

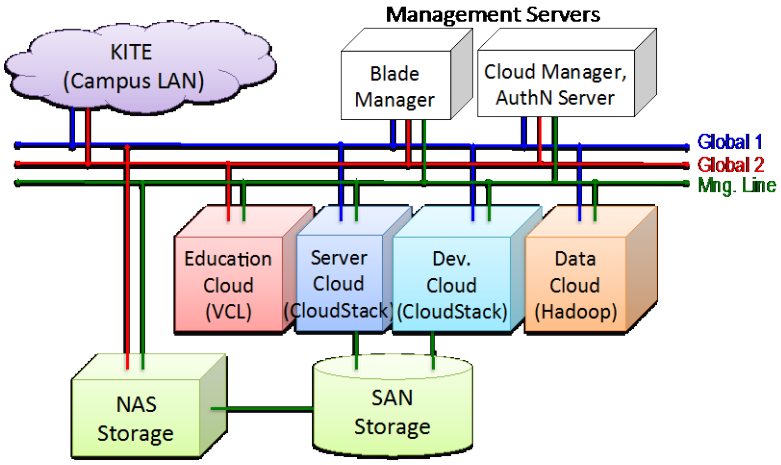

Figure 1. An outline of Qcloud.

Education cloud in Qcloud with VCL

This paper intends to practice the construction of search engines. We built it for practicing exercises on VCL for education. Figure 2 shows the architecture of the VCL based education cloud system. A user signs into the VCL system and submits a VM generation request to the system.

As shown in Figure 2, there is a file server in education cloud that stores user's files. Windows and Linux VMs are configured to automatically mount the user's private file space in the file server. The VM is deleted at the end of a lecture, because expire time is defined in

VCL. Files on the VM are also removed if the VM is deleted. Programing and data preparation may not finish in one day, and then files should be kept until the last day of the lecture. We installed a CIFS (common internet file system) user access module into the file server. Windows and MacOS support CIFS, so a user can access his/her file space anytime via the campus network without any preparation. The file server contributes to the improvement of user experiments.

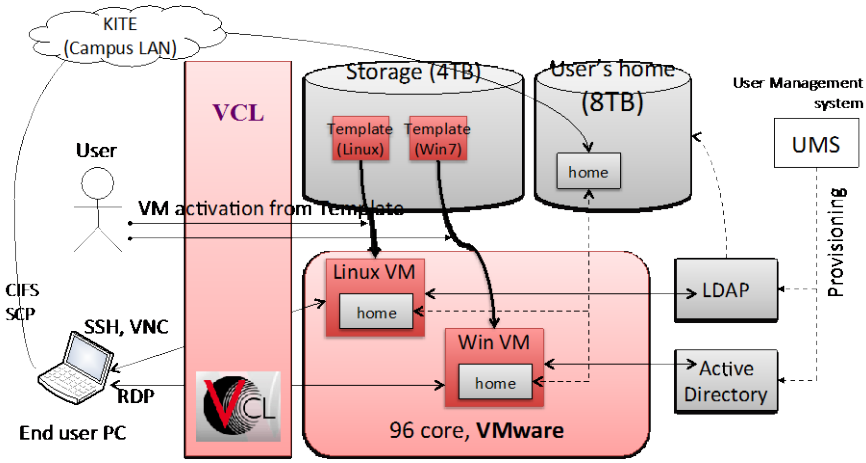

Figure 2. Education Cloud with VCL in Qcloud.

3. Web search engine exercise

Course description

We applied the use of the Qcloud system to the *Advanced Distributed Systems* course. Hirokawa, one of authors of this paper, has been teaching advanced distributed systems to graduate students since 2008. According to the syllabus of the course, lectures in this course are on the principle, theory, technologies, and applications of wide area network environments, and basic techniques of the web such as HTML, URI, and HTTP. Moreover, the course includes the processing and method for collecting data dispersed in the wide area, analyzing, unifying, and utilizing. Since the operation of the Qcloud system started in 2012, we decided to perform the construction exercise of a Web search engine from the lecture in the 2012 fiscal year.

Web search engine construction

In order to construct a Web search engine, it is necessary to collect and to preprocess data. An outline of preprocessing is shown in the left half of Figure 3. First is the collection of search data such as html files. Second is forming where low data files are written in various formats. The last is indexing, where an index file represents the relation of words and files.

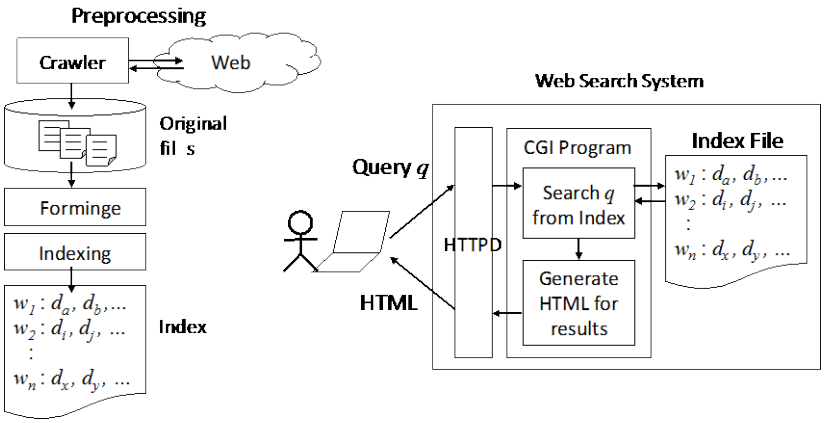

Figure 3. Preprocessing and Simple Web Search Engine.

An outline of a simple search system is shown in the right half of Figure 3. The system consists of a web server, search program, and index. In Figure 3, search programs are implemented as CGI (common gateway interface) programs. The user submits a query to the search system. The search CGI program gets the query and decomposes it into words. Then it scans for the words in the index. Finally, the program generates HTML for search results, and returns it to the user.

Customizing a Linux VM for web search engine construction

We made a Linux VM image for a web search engine. At the first stage, the VM was a very simple CentOS machine on which we installed Apache web server (ver.2.0), the morphological analysis tool *chasen* [11], and general-purpose associative retrieval engine *GETA* (ver.2) [10]. We also configured the VM appropriately for web search engine construction. The Apache web server starts when the machine starts-up. We configured the *httpd.conf* (Apache's access control file) to permit the execution of CGI programs on the website for users. We also changed the *iptables* file to permit HTTP web access (port 80) communication from university IP addresses.

The installation of software and configuration of server setting are required for practical system construction, however it is less related to the essence of the information retrieval. The essence of a search engine is the crawling of web data, the forming of collected data files, indexing of files, and the search algorithm. In the exercise course, we prepared an all-in-one packed VM image. Student's only need to start the VM to try exercises and can concentrate on the exercise.

Steps of web search engine construction

In the exercise in 2012, the lecture was carried out in the following steps.

1. Starting Linux VM on VCL. Basics of Linux operation, and basics of Perl language. (One day)
2. Word counting from original data and extraction of *WAM* (word-article matrix) file by Perl script. (One day)
3. Basics of GETA (Generic Engine for Transposable Association). Construction of a file search Perl program using GETA. (Two days)
4. Web CGI Perl programing. (One day)
5. Self exercise of web search engine construction using provided data files. (2days)
6. Presentation of exercise results. (One day)

In step 2, we offered 100 files of an online novel service metadata. Before the exercise we collected metadata files from syosetu.com, which is a Japanese online novel service. All metadata files were written in the YAML text format [12], which is a structured

data format, but very easy to understand the structure. So, these are suitable for novice word counting.

Before step 5 (the last exercise), we also offered 42,921 files. These files contain the outline data of articles, which were posted to IEICE society (the Institute of Electronics, Information and Communication Engineers), during 2004 to 2011. An article outline consists of a title, author(s), and abstract. The abstract part has long natural language (Japanese) sentences, and are difficult to word count. Then, we offered morphological analysis of sentence, and extracted words.

In step 5, we provided two data sets. Student selected either the metadata of online novels, or IEICE outline data, and built the search engine using the data. Although we allowed the collection of real web data, there was no student who collected data by themselves. Finally, students presented their search engine that he/she created. We indicated to students to show devised points, enhanced points, and invented points in the presentation.

4. Discussion

Questionnaires and results

In order to evaluate the course content, an easy questionnaire for attendance students was conducted. The survey items consist of the following seven items and got ten attendance students to evaluate each on a five-point scale. The results of mark evaluation is shown in Table 2. The right table in Table 2 is a modified version of the left, where the scores are subtracted by 3 for intelligibility.

a. Usability of VCL web interface
b. Length of VM (max:4 hours, min:1 hours.)
c. Usability of Linux VM
d. Usability of SSH terminal operation
e. Usability of Apache Web server
f. Usability of GETA
g. Usability of Perl Language and GETA Perl modules

Table 2: Questionnaire results (Left: original, Right: -3).

Answered student

Q	1	2	3	4	5	6	7	8	9	10	Ave.
a	4	4	5	3	5	4	4	5	5	4	4.3
b	2	3	4	1	4	2	4	5	4	4	3.3
c	4	3	5	5	4	4	3	5	5	5	4.3
d	4	3	4	5	4	4	4	5	5	5	4.3
e	3	4	5	3	4	4	3	2	5	4	3.7
f	3	4	3	4	4	3	3	5	4	3	3.6
g	3	4	4	5	4	3	4	4	4	4	3.9

Answered student

Q	1	2	3	4	5	6	7	8	9	10	Ave.
a	1	1	2	0	2	1	1	2	2	1	1.3
b	-1	0	1	-2	1	-1	1	2	1	1	0.3
c	1	0	2	2	1	1	0	2	2	2	1.3
d	1	0	1	2	1	1	1	2	2	2	1.3
e	0	1	2	0	1	1	0	-1	2	1	0.7
f	0	1	0	1	1	0	0	2	1	0	0.6
g	0	1	1	2	1	0	1	1	1	1	0.9

Discussion of questionnaire results

The questionnaire contains a portion in which students can describe their opinion freely. We observed through this portion that the attendance student's degree of satisfaction was high. The subject assigned to the students is to construct a search engine. They tackled the subject independently and enjoyed it very much.

The evaluation scores of (b), (e) and (f) were low in Table 2. The question (b) concerns to the usability of VCL system. Students wrote dissatisfaction to time restriction as their free opinion. This is because the time limit of virtual machine in the campus cloud system of Kyushu University is only 4 hours. This limit was set as default. But they can extend the limit to 8 hours. Unfortunately, they required much longer time for completing their task of constructing a search engine. They were forced to suspend their job during their

effort. The setting of a time limit was a problematic. However, students were absorbed in making their own search engine. Thus, the task was successful. Indeed, many students wrote that they were interested in creating a search engine.

The evaluation score of question (e) was low. It concerns with the usability of Web servers for CGI. This is because that the preparation of virtual machines was not enough. It was the first time to do this class in a cloud environment. It took several weeks until detailed settings of Apache and permission of related files were worked out. This delay caused troubles for students. Once the environment was settled, they did not have difficulty in the second half of the class. Unfortunately, the bad impression in the early times remained.

The reason for the low score for the question (f) comes from the complex of the exposition. They had to learn the basic notion of information retrieval, such as vector model of documents, term document matrix and indexing. Then, they had to learn how to use the tools to create indexes. Since there is a portion with a difficult understanding, it is thought that it had an impression that is hard to use.

Effectiveness of Cloud

We are satisfied with the effectiveness of the lecture and the exercise using virtual machine and cloud environment. The first advantage is the maintainability of computer environment. In case of customization a physical computer environment to fit a course, we must consider the affect of the customization on other courses. On the other hand, in case of customization of a VM on VCL, we do not have to care about customizations affecting other courses because the VM is only used by a course.

The second advantage of VM (or VCL) is that students can experience working as an administrator. In the exercise, students configured settings of the Apache's httpd.conf, and students satisfied these administration works, because they wanted to train computer system admonition for their work after graduation. It was difficult to train admonition in conventional physical computer environment, because it is too risky. A simple miss configuration easily causes malfunctioning for the whole system. On the other hand, on VCL, a learner can delete the VM when the VM's behavior became wrong because of misconfiguration.

5. Conclusion

This paper reported lessons taught in a search engine course for ICT major graduate students. They learned how to create their own search engine. The course was taught and run on a private cloud system. The structure of the cloud system and the contents of the course were described. The exercise, which actually builds a search engine on the virtual machine in cloud, was performed. The teachers collected the data and provided it as search targets.

As a result of summarizing the questionnaire from attendance students, it turned out that the attendance students had a very high degree of satisfaction. Analysis of the questionnaire results showed that a lecture is improvable by setting change of a VCL system. The advantages in the case of using the cloud system in education were described as well. We will expand the scale of a cloud style education system from now on, and consider utilizing it also for the education in faculties and graduate schools other than an information system. The content of system maintenance needs to be analyzed. We would like to consider offering the cloud system not only within our university, but also to domestic universities or other universities around the world.

References

Apache VCL, http://vcl.apache.org/ .

NCSU VCL, http://vcl.ncsu.edu/ .

Schaffer, H. E., Averitt, S. F., Hoit, M, I., Peeler, A., Sills, E. D., and Vouk, M. A. (2009). NCSU's Virtual Computing Lab: A Cloud Computing Solution, *Computer*, 42(7), 94-97.

Stein, S. R., Schaffer, H. E. (2010). Cloud with a Long Tail: The VCL in Support of Pedagogy, *EDUCAUSE Review*, 45(3), 10-11.
Apache CloudStack, http://cloudstack.apache.org/, (accessed at Aug. 8, 2013, UTC).
Hitachi (2012). Press release, http://www.hitachi.co.jp/Div/jkk/kyoiku/casestudy/kyushu2/casestudy1.html .
Ito, E., Kasahara, Y., Hori, Y., Inoue, K. (2013). A Study of VCL in Graduate School of ICT, IPSJ SIG Technical Report, 2013-CLE-9 (9), 1-6.
GETA, http://geta.ex.nii.ac.jp/, (accessed at Aug. 8, 2013, UTC).
Chasen, http://chasen.naist.jp/hiki/ChaSen/, (accessed at Aug. 8, 2013, UTC).
YAML in Wikipedia, http://en.wikipedia.org/wiki/YAML, (accessed at Aug. 8, 2013, UTC).