Gulliver’s Toss: Google’s Chronic Big Load to University Mail Server and Its Sudden Resolution

Kasahara, Yoshiaki
Kyushu University

Ito, Eisuke
Kyushu University

Fujimura, Naomi
Kyushu University

https://hdl.handle.net/2324/4796096

出版情報:2013-11-03. Association for Computing Machinery: ACM
バージョン:
権利関係:Copyright © 2013 ACM
Gulliver's Toss: Google's Chronic Big Load to University Mail Server and Its Sudden Resolution

Yoshiaki Kasahara
Kyushu University
6-10-1 Hakozaki, Higashi-ku
Fukuoka 812-8581, Japan
+81 92 642 2297
kasahara.yoshiaki.820@m.kyushu-u.ac.jp

Eisuke Ito
Kyushu University
6-10-1 Hakozaki, Higashi-ku
Fukuoka 812-8581, Japan
+81 92 642 4037
ito.eisuke.523@m.kyushu-u.ac.jp

Naomi Fujimura
Kyushu University
4-9-1 Shiobaru, Minami-ku
Fukuoka, Japan
+81 92 553 4434
fujimura.naomi.274@m.kyushu-u.ac.jp

ABSTRACT
Traditionally, Kyushu University has been providing email service internally using its own domain name for staff members and students of the university. Around January 2012, we noticed that the high load of the university authentication server, and we realized that one of causes was the access from the mail server for students (called Student Primary Mail Service). Detailed analysis showed that there was chronic big load produced by Gmail's Mail Fetcher, especially toward nonexistent accounts removed due to graduation. In this paper, we explain the situation and reasons of the big load induced by Google, its possible countermeasures, and its sudden resolution by Google's silent change.

Categories and Subject Descriptors
H.4.3 [Information Systems Applications]: Communication Applications – Electronic mail

Keywords
University mail service; On-premises service; Deleted accounts; User authentication.

1. INTRODUCTION
Information and communication service is indispensable for education and research activities in universities. Among various services, electronic mail is one of the fundamental services even before the Internet had been widely available in the world. From the perspective of equipment, mail service offered by a university can be categorized into two kinds. One is an in-house (on-premises) service operated by the university itself, and the other is SaaS (Software as a Service) type service operated by other parties. SaaS-type email services such as Google’s Gmail, Yahoo!

Mail, and Windows Live@edu (now known as Office365) became popular around 2007.

We aren’t going to discuss which is superior here, but for various reasons, Kyushu University has been providing on-premises email services using its own domain name for our university staff members and students for several years. Email service for students was started in 1995 (the current incarnation is called “Kyushu University Student Primary Mail System”) [1][2], and another service for university staff members was started in July 2009 (called “Kyushu University Primary Mail Service”) [3].

Various network services (including these email services) inside our university utilize the university-wide authentication service, and we have been analyzing the load of the authentication servers [4]. We realized that requests from the university-wide wireless LAN service and the student email service was enormous and dominated more than 95% of the whole requests to the authentication servers. In this paper, we describe more detailed analysis of the student email server which revealed that there were a lot of POP3 accesses from Google’s network ranges to the email server. In fact, 57% out of all the POP3/IMAP4 requests to the student mail server came from Google. In addition to that, 65% of Google’s accesses were toward non-existent accounts (mostly removed by graduation).

In this paper, we describe the analysis of the heavy load induced by Google to the student email server and its consequence. The rest of this paper is organized as follows. In Section 2 we describe email services in Kyushu University. In Section 3 we describe the situation of the student email server’s load by analyzing log files of the student email server. In Section 4 we discuss possible countermeasures against Google’s accesses. In Section 5, we report the sudden resolution of the problem. Finally, we present our conclusions in Section 6.

2. PRIMARY MAIL SERVICES
First of all, we explain about university-wide mail services provided by Kyushu University Information Infrastructure Initiative.

2.1 Number of Users
The main members of our university are students, faculty, and staff members. Most of students are undergraduate and graduate curricular students. There are also some non-curricular students such as research students and special register students. In addition to students, faculty, and staff members, there are other members
Table 1. The number of IDs in Kyushu University (Mar 2012)

<table>
<thead>
<tr>
<th>Role</th>
<th>Total No. of IDs (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curricular students</td>
<td>19,000</td>
</tr>
<tr>
<td>Non-curricular students</td>
<td>500</td>
</tr>
<tr>
<td>Faculty and staff members</td>
<td>9,000</td>
</tr>
<tr>
<td>Temporary staff etc.</td>
<td>800</td>
</tr>
<tr>
<td>Total</td>
<td>29,300</td>
</tr>
</tbody>
</table>

with various roles such as research fellows, temporary staff, visiting researchers, and so on. Table 1 shows the approximate number of IDs issued for the university-wide authentication service. The number also represents how many people can use both “Primary Mail Services”.

The approximate number of ID turnover per year is 6,000 for students. It includes about 3,000 undergraduate students and 3,000 graduate students. In other words, every year 6,000 accounts are both created and deleted on Student Primary Mail Service.

On the other hand, turnover of faculty and staff members is about 1,000 per year and mostly happen in March and April (1,000 retired or moved out of the university in March and 1,000 added in April). Due to that, every year 1,000 accounts are both created and deleted on Primary Mail Service.

2.2 System Overview

Because Primary Mail Service (for faculty and staff members) have been using separated authentication servers (due to some service requirements) and irrelevant to the discussion in this paper, we only explain the system overview of Student Primary Mail Service.

Figure 1 shows the system overview of Student Primary Mail Service. The system employs user authentication through SMTP-AUTH for email submission to avoid third party relaying of spam messages. Users retrieve their email messages through standard POP3/IMAP4 over SSL/TLS with user authentication. Users won’t do shell login to the server directly.

The system had been built using open source software such as Postfix and Dovecot on top of general purpose IA servers to reduce the system cost. To support relatively large number of students, we needed to concentrate budgets to the spool storage.

3. SYSTEM LOAD INDUCED BY GOOGLE

In this section, we will explain how we noticed chronic load induced by Google, and its analysis in details. Actually the discussion from here to Section 4 was applicable as of September 2012 and already resolved now as described in Section 5, but we want to present them here for references.

3.1 Outline

The university-wide authentication system included LDAP servers for processing user authentication and account information retrieval. They were used by not only “Primary” email services but also various information services in our university. From around January 2012, intermittent outage of authentication service started to be observed, causing user authentication failures of various services. We investigated the reason and realized that the LDAP servers were simply overloaded by too many requests and

Table 2. Access statistics of Student Primary Mail (June 2012)

<table>
<thead>
<tr>
<th></th>
<th>POP/IMAP accesses per day (ave.)</th>
<th>Unique IDs accessed per day (ave.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>234,843 (100.0%)</td>
<td>19,996 (100.0%)</td>
</tr>
<tr>
<td>Authentication successful</td>
<td></td>
<td></td>
</tr>
<tr>
<td>from Google</td>
<td>113,491 (48.3%)</td>
<td>6,435 (32.2%)</td>
</tr>
<tr>
<td>from campus LAN</td>
<td>33,835 (14.4%)</td>
<td>2,732 (13.7%)</td>
</tr>
<tr>
<td>Authentication failure by incorrect password</td>
<td>19,081 (8.1%)</td>
<td>3,432 (17.2%)</td>
</tr>
<tr>
<td>from Google</td>
<td>19,081 (8.1%)</td>
<td>3,432 (17.2%)</td>
</tr>
<tr>
<td>from campus LAN</td>
<td>6,483 (2.8%)</td>
<td>689 (3.4%)</td>
</tr>
<tr>
<td>Accesses to non-existent IDs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>from Google</td>
<td>90,773 (38.7%)</td>
<td>3,989 -</td>
</tr>
<tr>
<td>from campus LAN</td>
<td>86,495 (36.8%)</td>
<td>3,567 -</td>
</tr>
<tr>
<td>Other errors (no ID provided etc.)</td>
<td>1,266 (0.5%)</td>
<td>190 -</td>
</tr>
</tbody>
</table>
failed to process them.

The two largest users of these LDAP servers were university-wide wireless LAN service using 802.1x authentication, and Student Primary Mail Service (Primary Mail Service for staff members had already migrated to their own LDAP servers). Especially, Student Primary Mail Service employed the LDAP bind operation to authenticate users so we considered that it must be more demanding to the LDAP servers.

By investigating the log files of the mail server, we found that there was unexpectedly large number of POP3 authentication requests from Google’s network addresses. Actually we also realized during investigation that these LDAP servers were not properly tuned because initially they were prepared for web services and not expected to serve such a huge number of requests. If the LDAP servers had been optimized enough, probably we wouldn’t notice the load from Google.

3.2 Analysis

To understand what was going on, first we investigated log files of the mail server from Dovecot POP3/IMAP4 software and PAM system used for user authentication backend of Postfix SMTP software for one month (June 2012). We gathered the number of login attempts and number of unique IDs used. We didn’t distinguish POP3 and IMAP4 accesses because the authentication process of them was the same (using PAM). During analysis, we noticed that there seemed to be many accesses from Google’s network addresses, so we also distinguished accesses from Google and campus LAN to compare each other. Table 2 shows the results.

From Table 2, about one-third of total users (6,435 out of 19,996) accessed the mail server using POP3/IMAP4 during the period, and about 42% (2,732 out of 6,435) of users were from Google’s network addresses, probably retrieving messages through Gmail. Also it was notable that most of accesses toward non-existent accounts were also from Google. By the access counts, 39% (90,773 out of 234,843) of accesses were toward non-existent, invalid accounts, and its 95% were from Google. That means that if these invalid accesses from Google disappeared, about 37% of the total load of login authentication processes is gone. In total, 57% (132,752 out of 234,843) of all POP3/IMAP4 accesses were from Google, and 65% (86,495 out of 132,752) of all accesses from Google were toward non-existent accounts. Actually we didn’t expect this large number of accesses toward non-existent accounts from Google at all. Google’s accesses toward non-existent accounts seemed irregular compared to accesses from campus LAN.

Next, we analyzed chronological trends of POP3/IMAP4 accesses from February 2011 to May 2012. Figure 2 shows the number of login attempts via POP3/IMAP4 and the number of unique IDs used per day during the time period. These graphs show both the number of accesses toward non-existent accounts and the number of unique (non-existent) accounts were almost stable during the time period, and there was a sudden gap in the middle of May 2012. The gap represented the deletion of graduated students’ accounts, and about 1,200 (out of 6,000 deleted) of now-invalid accounts were added as invalid accesses. The graduation happened in the end of March, but the account deletion was committed in the middle of May due to operational reasons. As mentioned before, most of these accesses were from Google. It seemed that many graduated students didn’t change their Gmail settings, and Google didn’t care about authentication failures too.

3.3 Google’s Mail Fetcher Service

We couldn’t tell exactly what software was used to generate these accesses from Google by our log files, but we believe that Google’s Mail Fetcher service was used. Google’s help document says that Mail Fetcher can be used “to centralize all your email from different accounts into one Gmail account.” Mail Fetcher uses POP3 protocol to collect email messages from other (possibly multiple) email servers and stores them in the mailbox of Gmail.

We guessed some possible reasons why many students were using Mail Fetcher. One reason was that we explicitly informed students how to move their messages from our service to another email service. Until 2009, we provided webmail as a part of student email service, but it was abolished when the entire educational system was replaced in 2009 due to budget restriction. The vendor support of the previous webmail software had already ended, and we couldn’t prepare a good alternative with sufficient server power to support all the students within our limited budget. Actually, the previous webmail was under-powered and there were a lot of complaints. So we considered that we should provide alternative methods to access their messages conveniently. Especially for Gmail, we also provided how to fetch their messages through Mail Fetcher. In addition to that, in some departments, there was a hands-on session of obtaining Google account and configuring Mail Fetcher during introductory lecture of information technology. So basically we encouraged students to use Mail Fetcher (without realizing the consequences).
Usually using Mail Fetcher is believed to be preferable over simple email forwarding, because email forwarding has bad effects to spam filtering. For example, when a user configures to forward all the messages to another mail system, spam messages are also forwarded. If a lot of users do the same thing, the origin of forwarded messages may be mistakenly flagged as a spam source (it depends on the way of spam detection, though) and causes lower reputation and even in some blacklists. Also announcement messages from the university to all the students may be treated as spam because of sending the same content to many accounts of the destination server. It is common to share accounts and non-existent accounts. More than 90% of accounts were accessed once per hour (24 times in a day) and 48 and 72). That means that Google controlled the access frequency by using the last record of message arrival rate when the account had been still accessible.

4. IMPACT AND COUNTERMEASURE
In this section, we discuss the impact of accesses from Google, and possible ways to circumvent it.

4.1 Impact of Accesses from Google
By discussion in Section 2.1 and 3.2, we realized that about 6,000 accounts were replaced every year and about 1,200 (~one-fifth of 6,000) accounts were accessed continuously even after removed. The current incarnation of Student Primary Mail Service began its service in 2011, but it had inherited the previous server’s FQDN for mail client migration from the previous system (in service from 2009). From 2009 these invalid accesses to deleted account were accumulated, resulting about 3,600 (~3 years’ worth) deleted accounts still accessed by Google.

To estimate the scale of the impact of accesses from Google, we checked the total number of university students in Japan. According to the School Basic Survey of the Ministry of Education, Culture, Sports, Science and Technology (in fiscal year 2010), there were 2.9M university students including national, public and private universities in Japan [6]. If all the students had email accounts provided by their universities, and one-third of students were replaced every year, then about 1M accounts are replaced every year. In Kyushu University, about 20% of them were considered using Google’s Mail Fetcher, but it must be over the average. Let’s assume 10%, and every year about 100,000 accounts were accumulated as invalid accesses to deleted accounts from Google. The total number of Gmail accounts was 350M in 2012 [7], so only 100,000 invalid accounts per year should be negligible for Google.

Such accesses from Google continue to give useless loads over various mail services all over the world. Especially mail services in universities have higher rate of account turnover than in enterprises, so adverse effect must be greater. Also, when designing services for students, it is usually assumed that the number of students is mostly constant and the budget is limited, so it is likely that constant increase of such load interfere the normal operation of the system.

4.2 Possible Countermeasures
We discuss the following methods as countermeasures for increasing access from Google.

1) Filter out POP accesses from Google
2) Ask users of Mail Fetcher to change their configuration
3) Optimize mail system to reduce load
4) Ask Google to change its access policy

4.2.1 Filter out POP accesses from Google
This is the most instantaneous and effective measure to remove the load induced by Google. It is pretty easy to implement such a filter, but it also greatly reduces users’ convenience, so it should be considered a last resort. There was one case of this method actually implemented in Hokkaido University of Education because the load from Google was high enough to hinder ordinary users’ accesses.
4.2.2 Ask users to change their configuration

This is a method which tries to remove loads by reducing the number of Mail Fetcher users. When users, whose Mail Fetcher setting has problems, are still enrolled in the university, we can contact them (via email or bulletin boards) and make them fix their configuration. But actually most of the load we want to reduce is caused by deleted accounts of graduated students, and it is hard to contact them already. Also we cannot contact the original Gmail account because no information is provided from Mail Fetcher to the destination server who is using the Mail Fetcher.

One possibility is to revive invalid accounts temporarily and put a message into the mailboxes which directs the users to remove Mail Fetcher setting, and let Mail Fetcher fetch the message to the corresponding Gmail account. It is not a trivial task, and also it has limited effects because some users might abandon Gmail accounts and don’t read messages anymore.

4.2.3 Optimize mail system

Another way is to reduce the overall system load itself by changing configuration of servers. For example Dovecot has an ability to cache authentication information (such as IDs and passwords) to reduce load of the authentication server. But enabling it should be carefully considered because such a cache always has a problem of cache inconsistencies and synchronization, especially when a user changes his/her password. Another method is to route accesses from Google into other servers using a load balancer. But the load from Google will accumulate every year, so these methods may have limited effects. It is hard to continue investing in system for deleted accounts.

4.2.4 Ask Google to change its policy

The fundamental problem is that Google doesn’t care about unsuccessful logins and continues the accesses indefinitely. It should be easy for Google to improve Mail Fetcher to solve the problem. For example, it is common to implement back-off (extends intervals between failed logins) and time-out (gives up accessing inaccessible accounts after several trials) processing in case of login failures. When time-out occurred and Mail Fetcher was disabled, the event should be reported to the original Gmail account by putting a message into the mailbox. If the user is actively using the account, he/she will notice the warning and change the configuration. If the account has been abandoned, it is also fine because Mail Fetcher stays disabled anyway.

A problem is that there is no clear way to inform Google about such a problem. Feature requests to Google supposed to be posted into Google Product Forums. We found that there was an article in the forum about a similar request posted in August 2010, but there was no response and no improvement as of September 2012.

5. RESOLUTION

After we realized about the problem with Google’s Mail Fetcher, we decided to present the issue in a domestic workshop to gather information [5]. We also considered to bring it to an international conference such as ACM SIGUCCS to draw attention and gather information about the status in other organizations outside Japan. But Google actually took the initiative.

Just one day before the session of the domestic workshop in September 2012, we realized that the number of authentication failures had suddenly decreased in the beginning of September. Figure 4 shows the number of login attempts and accessed unique IDs per day (including Google’s change).

The graphs show that the number of login attempts suddenly decreased to about one-tenth, while the number of unique accounts is not changed. It seemed that Google silently decided to stop accessing inaccessible accounts every hour and only tried once or twice in a day. By this change, the authentication failure from Google became almost negligible and the issue was resolved by Google’s own discretion. We couldn’t find any announcement or blog entries by Google about that change.
Figure 5 is the same graph as Figure 3 after Google’s change (from September 11th to 17th, 2012). It clearly shows that Google started to treat inaccessible accounts differently from accessible accounts. More than 95% of non-existent accounts were accessed only once a day, but some accounts were still accessed twice to several times a day.

6. CONCLUSION
In this paper, we described details of a chronic load induced by Google’s Mail Fetcher, its possible countermeasures, and its sudden resolution. We analyzed log files of Student Primary Mail service to show the load status of these servers. As a result, we realized that Google caused a chronic load to our university mail server. We also discussed about the impact of Google’s access and possible ways to mitigate it.

This particular issue presented in this paper has been resolved by itself without our actual intervention. That means that such an issue may happen again when these giant service providers such as Google, Microsoft, and Amazon decided to introduce a new service or modify their existing services. It must be done by their own goodness, but it is like a DoS (denial of service) attack for our small installment of services when their pinkie finger comes toward us. It is hard to predict and prepare for such an event, but at least we should take them into consideration when designing our system, and be observable to the status of our own system.

Sometimes we should act quickly against their moves to protect our own services and our users.

7. ACKNOWLEDGMENTS
Our thanks to all students who are using our services, and staff members of the Primary Mail Service working group and the Authentication Infrastructure working group to develop and maintain these systems in Information Infrastructure Initiative of Kyushu University.

8. REFERENCES