OSG Certificate Usage Analysis

Von Welch

DRAFT

April 30,2012

# Data

Data in this document is based on data obtained from two sources:

1. A ldif dump received from ESnet on certificate issuances by the DOE Grids PKI
2. Data extracted from OIM regarding OIM-registered OSG users

Please see the appendices for details on this data.

All data files are archived at:  
gocbox.grid.iu.edu:/usr/local/vwelch.

All analysis scripts are archived at:  
https://vdt.cs.wisc.edu/svn/software/doe-cert-analyze/

# Known Data Limitations in Estimating OSG Need

There following are known limitations of the data received from ESnet:

1. Non-OSG Certificates: The data includes all DOE Grids PKI certificates, not just OSG. The analysis in this document ignores this under the assumption that OSG (and other VOs OSG plans to support) is the vast majority of DOE PKI usage.
2. Erroneous Issuances: Not all certificateRecords are truly issued certificates delivered to users. During the analysis it was noted that one OSG staff member had 11 certificateRecords and was contacted about why this might be the case. He indicated he had problems renewing and had to try multiple times. Further analysis did indicate that most of the 11 certificateRecords were very proximate in time, which indicates that while the system logged multiple issuances, they were not delivered to the user. In 2011, 43 users had 4 or more certificate issuances indicating this may be a common problem.
3. Multiple Host Issuances: A number of hosts are seen to have many certificateRecords (in 2011, 2 had >60, 2 others > 30, 3 others > 20, and 11 others > 10; 231 had 4 or more issuances). The reason has not been determined. Reasonable assumptions would seem to be errors as in the previous limitation or, given issuances are free, it was just expedient for some system administration reason to request new certificates instead of maintaining current ones (e.g., during OS reinstallations).
4. Partial 2012 data: All data for 2012 is through April 18, 2012. On cursory observation, 2012 data seems to be consistent with previous years.

The following are known limitations of the data received from OIM:

1. The data represents only a fraction of total OSG users: The OIM data only contains entries for approximately 630 users. This is only approximately a quarter of OSG’s user base.

# Analysis of Certificate Issuances

|  |  |  |
| --- | --- | --- |
| Year | # Records  (ldif-analyze.py count) | # Unique Subjects (ldif-analyze.py count-subjects) |
| 2003 | 1340 | 1273 |
| 2004 | 2115 | 1979 |
| 2005 | 4104 | 3643 |
| 2006 | 5132 | 4632 |
| 2007 | 7985 | 6438 |
| 2008 | 9148 | 7914 |
| 2009 | 10741 | 8935 |
| 2010 | 11784 | 10413 |
| 2011 | 12514 | 10687 |
| 2012 | 3630 | 3112 |
| 2009-2012 combined | 38669 | 17246 |
| All years combined | 68495 | 26523 |

Table : Number of certificateRecords and unique Subject names by year.

The difference between the columns is number of certificates issued versus number of unique subject names. This table seems to show a 10-20% difference. Reasons for this difference could include:

* Reissuance – individuals requesting a certificate multiple times in a given calendar year for renewal or lost key.
* Erroneous issuances – as described in Section 2.

The values in these two columns roughly bound the actual number of certificate issuances OSG needs, with the second column being the actual need if every user was to only receive one certificate per year (actually it would be slightly less given some users have multiple certificates).

## User Certificates

|  |  |  |  |
| --- | --- | --- | --- |
| Year | # User Certificates (ldif-analyze.py count-users) | # Unique User Subjects (ldif-analyze.py count-user-subjects) | # Unique Legal Names (ldif-analyze.py count-names) |
| 2003 | 517 | 493 | 431 |
| 2004 | 885 | 829 | 701 |
| 2005 | 1216 | 1085 | 985 |
| 2006 | 1522 | 1351 | 1262 |
| 2007 | 2252 | 1987 | 1817 |
| 2008 | 2634 | 2178 | 1956 |
| 2009 | 2764 | 2388 | 2171 |
| 2010 | 3182 | 2653 | 2417 |
| 2011 | 3129 | 2562 | 2328 |
| 2012[[1]](#footnote-2) | 925 | 781 | 733 |
| 2009-2012 combined | 10000[[2]](#footnote-3) | 5206 | 3957 |
| All years combined | 19026 | 9376 | 6215 |

The difference between the first and second columns indicates the number of multiple issuances (some perhaps erroneously as described in Section 2). The difference between the second and third column indicates the number of people who may have multiple certificates issued to them (it also includes people who share a legal name, so this is not fully accurate).

**Observation**: OSG has about a 20% re-issuance rate per year for user certificates.

**Observation**: OSG has < 10% of its community that has multiple certificates with different subject names. It has been noted that ESnet staff, who obtain certificates from the DOE Grids PKI, account for some portion of this as they obtain multiple certificates (6-10 have been observed) regularly for different devices.

**Observation:** The OSG user certificate usage seems to have leveled off over the past couple years at approximately 3200 issuances for 2600 subject names.

## Host Certificates

|  |  |  |  |
| --- | --- | --- | --- |
| Year | # Host Certificates (ldif-analyze.py count-hosts) | # Unique Hostnames  (ldif-analyze.py count-hosts) | # 2nd-level domains (ldif-analyze.py count-domains) |
| 2003 | 822 | 713 | 64 |
| 2004 | 1227 | 1047 | 94 |
| 2005 | 2882 | 2266 | 109 |
| 2006 | 3605 | 2982 | 131 |
| 2007 | 5733 | 3989 | 138 |
| 2008 | 6513 | 5169 | 131 |
| 2009 | 7975 | 5797 | 135 |
| 2010 | 8601 | 6849 | 129 |
| 2011 | 9385 | 7086 | 136 |
| 2012[[3]](#footnote-4) | 2705 | 2044 | 79 |
| 2009-2012 combined | 28666 | 9996 | 191 |
| All years combined | 49449 | 13920 | 284 |

The difference between the first and second columns indicates the number of re-issuances (some perhaps erroneously as described in Section 2).

The third column shows the number of second-level domains (e.g., fnal.gov) that appear in the certificates. Note that there are a number of host certificates issued each year without valid domains (e.g., 171 in 2008, 148 in 2009, 480 in 2010, 651 in 2011, 255 in 2012).

**Observation:** OSG has about 25-35% re-issuance rate per year for host certificates.

**Observation**: The number of domains has been relatively stable since 2006. However, the data indicates some amount of churn in the domains as there have been 191 different domain names represented since 2009.

**Observation:** OSG’s host certificate usage seems to be steadily growing at 10% per year.

## Foreign Host Certificates

The new OSG PKI will primarily[[4]](#footnote-5) be serving users in the U.S. The command “count-non-osg-domains” attempts to count host certificates issued to hosts outside of the U.S. Running it on ldif data back to the beginning of 2011 (“cat 2011.ldif 2012.ldif | ldif-analyze.py count-non-osg-domains”) resulted in 298 certificates issued across 48 domains. The exact semantics of the command were to look for hosts out side of the following domains: .edu, .org, .gov, .net, and .us.

This list contained an erroneous entry in the form of a ligo host with a bad hostname (“ldas-cit.ligo.caltech”) and two host certificates issued for a FutureGrid host with a dnsdynamic.com hostname.

The list of domains can be found in Appendix C.

Visual sorting of the list indicated most of the host certificates were issued to hosts in Korea, Brazil, Columbia, China and Japan, with a few certificates being issued to hosts at CERN and in Germany, France and the UK. No commercial certificates (.com) were noted.

Administrators of the various host certificates can be determined by using a web browser to visit <https://myosg.grid.iu.edu/about> and entering the domain name.

**Observation:** OSG’s non-U.S. host certificates are approximately 5% of its total host certificate issuances.

## Foreign User Certificates

The new OSG PKI will primarily be serving users in the U.S. Since there is no field in the certificates that indicates a user’s geographic location or nationality, the OIM user data was analyzed for obvious foreign email addresses: “cat OSG-user-data-from-oim.csv | oim-csv-parse-non-us-emails.py”

Of the 629 records, 424 were from the DOE Grids PKI, and only 25 were obviously non-U.S. (defined as not being from one of the following top-level domains: .edu, .org, .gov, .net, .us, .com). The domains of these email addresses are in Appendix D (email addresses are not included in this document for privacy).

Not included in this number were 24 email addresses from .com domains, 17 of which were gmail.com. These addresses are ambiguous as to their country.

**Observation:** Assuming linear interpolation from OIM-registered users to OSG’s total user population, OSG’s non-U.S. user certificates is approximately 5% of its total user certificate issuances.

# OSG Certificate Usage Estimates for 2013

This section present several possible estimates for OSG certificate usage for 2013, as an approximate timeframe for the first year of the new OSG PKI.

The partial data for 2012 appears to be consistent for the 3 years prior and this analysis assumes as much.

## Conservative Estimate

The most conservative estimate takes the number of certificateRecords from the DOE Grids PKI, interpolates that for OSG growth (10% for host certificates, flat for user certificates) and then adds approximately a 10% margin for error (10% + enough to round up to the nearest round number).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Previous Year | | With Growth | | With Margin of Error (~10%) | |
| User | Host | User (0%) | Host (10%) | User | Host |
| 2012 | 3129 | 9385 | 3129 | 10324 | 3500 | 11500 |
| 2013 | 3129 | 10324 | 3129 | 11356 | 3500 | 12500 |

## Optimistic Estimate

In the most optimistic case, we take the number of subjects that OSG has issued to, assume OSG will only issue one certificate per subject per year, account for growth, and take no margin for error.

This is unrealistic as it assumes no re-issuance for key loss or other error, but serves as a lower bound.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | Previous Year | | With Growth | |
| User | Host | User (0%) | Host (10%) |
| 2012 | 2562 | 7086 | 2562 | 7795 |
| 2013 | 2562 | 7795 | 2562 | 8574 |

## Moderate Estimate

In this case, we assume half of re-issuances are in error and those will be eliminated, we then add an approximately 5% margin for error (5% + enough to round up to the nearest round number).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Previous Year | | With Growth | | With Margin of Error (~5%) | |
| User | Host | User (0%) | Host (10%) | User | Host |
| 2012 | 2845 | 8235 | 2845 | 9059 | 3000 | 9600 |
| 2013 | 2845 | 9059 | 2845 | 9964 | 3000 | 10500 |

DOE Grids Certificate Data

On April 19, 2012 OSG received from ESnet an ldif file containing a dump of the DOE Grids PKI database of certificate issuance data:

$ ls -l 2012\_04\_18\_010121.ldif

-rw-r--r--@ 1 vwelch staff 863050777 Apr 19 14:55 2012\_04\_18\_010121.ldif

$ md5 2012\_04\_18\_010121.ldif

MD5 (2012\_04\_18\_010121.ldif) = 79e1cca378cb2bb74a8e94d9393a4d63

Many of the records were not regarding certificate issuances. Filtering the ldif records for those that have an objectClass of “certificateRecord”[[5]](#footnote-6) resulted in the following:

$ ls -l DOE-certificateRecords.ldif

-rw-r--r-- 1 vwelch staff 190352247 Apr 19 16:13 DOE-certificateRecords.ldif

$ md5 DOE-certificateRecords.ldif

MD5 (DOE-certificateRecords.ldif) = 62046cca01e6d6ece8231f322018fc3f

The file was then split by year:

$ for year in 2003 2004 2005 2006 2007 2008 2009 2010 2012 ; do cat DOE-certificateRecords.ldif | ldif-analyze.py filter-by-year ${year} > ${year}.ldif ; done

$ ls -l 20\*.ldif

-rw-r--r-- 1 vwelch staff 3255593 Apr 23 21:17 2003.ldif

-rw-r--r-- 1 vwelch staff 5161737 Apr 23 21:17 2004.ldif

-rw-r--r-- 1 vwelch staff 9983649 Apr 23 21:18 2005.ldif

-rw-r--r-- 1 vwelch staff 13371053 Apr 23 21:19 2006.ldif

-rw-r--r-- 1 vwelch staff 22268205 Apr 23 21:19 2007.ldif

-rw-r--r-- 1 vwelch staff 25630822 Apr 23 21:20 2008.ldif

-rw-r--r-- 1 vwelch staff 30489471 Apr 23 21:21 2009.ldif

-rw-r--r-- 1 vwelch staff 33595163 Apr 23 21:22 2010.ldif

-rw-r--r-- 1 vwelch staff 35872793 Apr 23 20:27 2011.ldif

-rw-r--r-- 1 vwelch staff 10718841 Apr 23 21:22 2012.ldif

All subsequent analysis described in this document was performed on these per-year files.

These data files are archived at:  
gocbox.grid.iu.edu:/usr/local/vwelch.

The analysis scripts are archived at:  
https://vdt.cs.wisc.edu/svn/software/doe-cert-ldif-analyze/

OIM User Data

OIM user data was obtained by visiting the following URL with a web browser and selecting CSV to download the user data in CSV format.

<https://myosg.grid.iu.edu/miscuser?count_sg_1&count_active=on&count_enabled=on&datasource=user>

Data size and has:

$ ls -l OSG-user-data-from-oim.csv

-rw-r--r--@ 1 vwelch staff 156153 May 3 14:31 OSG-user-data-from-oim.csv

$ md5 OSG-user-data-from-oim.csv

MD5 (OSG-user-data-from-oim.csv) = 1d95f54dfdc93384616a4ae9f70e88dd

These data files are archived at:  
gocbox.grid.iu.edu:/usr/local/vwelch

Foreign Host Domains

The list of non-U.S. host domains as discussed in Section 3.3 follows.

Total Non-OSG Domains: 48

Total Certificates: 298

knu.ac.kr: 52

hepgrid.uerj.br: 33

grid.unesp.br: 28

javeriana.edu.co: 21

sprace.org.br: 18

fis.cinvestav.mx: 12

phys.uregina.ca: 10

lcca.usp.br: 9

farm.particle.cz: 6

ucatolica.edu.co: 6

javerianacali.edu.co: 6

manchester.ac.uk: 6

upb.edu.co: 6

ustc.edu.cn: 5

udistrital.edu.co: 4

ucc.edu.co: 4

unitecnologica.edu.co: 4

ucaldas.edu.co: 4

uao.edu.co: 4

uniandes.edu.co: 4

autonoma.edu.co: 4

uautonoma.edu.co: 4

atum-sw01.cern.ch: 3

uac.edu.co: 3

mars04.cern.ch: 2

ccjbox5x.riken.jp: 2

mars01.cern.ch: 2

venus.cern.ch: 2

utp.edu.co: 2

kvm01.cern.ch: 2

ccjbox7x.riken.jp: 2

uniatlantico.edu.co: 2

ccjbox6x.riken.jp: 2

udem.edu.co: 2

uis.edu.co: 2

mars03.cern.ch: 2

riken.go.jp: 2

mars05.cern.ch: 2

osg.cnic.cn: 2

ccsvli50.in2p3.fr: 2

ttu-itb-futuregrid1.dnsdynamic.com: 2

ccjbox8x.riken.jp: 2

spserv02.sprace.br: 1

ccsvli02.in2p3.fr: 1

unab.edu.co: 1

ldas-cit.ligo.caltech: 1

d0-kit.gridka.de: 1

ncc.unesp.br: 1

OIM User Email Domains

The list of non-U.S. email domains associated with OIM-registered certificates as discussed in Section 3.4 follows

Total records: 629

DOE Grids records: 424

Non-US users: 35

Errors: 9

cern.ch: 18

ncc.unesp.br: 4

uerj.br: 2

ucaldas.edu.co: 1

usp.br: 1

physics.gla.ac.uk: 1

pi.infn.it: 1

mail.cern.ch: 1

correounivalle.edu.co: 1

unab.edu.co: 1

uniandes.edu.co: 1

utbvirtual.edu.co: 1

ift.unesp.br: 1

uac.edu.co: 1

1. Partial year data through April 18, 2012. [↑](#footnote-ref-2)
2. Yes, it’s really 10,000. [↑](#footnote-ref-3)
3. Partial year data through April 18, 2012. [↑](#footnote-ref-4)
4. The exact usage policy is still being decided on at the time of this writing. [↑](#footnote-ref-5)
5. Using ldif-output-certificateRecords.py [↑](#footnote-ref-6)