Systems Engineering Update
OARC 35, The Ether

Matthew Pounsett

2021/04/20

Contents

Contents 1

1 Introduction 2

2 OARC Services Overview 2
- 2.1 Data Archiving . . . 2
- 2.2 File Servers and Storage . . . 3
- 2.3 Data Analysis Servers . . . 3

3 System and Service Status 4
- 3.1 General Condition . . . 4
- 3.2 Authoritative DNS Improvements . . . 4
- 3.3 General Cleanup . . . 5
- 3.4 Fully Automated (OS) Installation . . . 5
- 3.5 Analysis Servers . . . 5
- 3.6 Discontinuing the Shared DSC Platform . . . 6
- 3.7 Networks, Routing, and Routers . . . 6
- 3.8 File Servers . . . 7
- 3.9 Day in the Life Dataset . . . 8

1 Introduction

The last six months of Systems/Network Engineering has been largely about maintenance, cleanup, and responding to the occasional emergency, more than working on some of our long term projects.

I'm still finding work on hardware in Fremont and Ottawa to be challenging, as we continue to rely on varying types and skill levels of remote hands. We're still hunting for people to fill short and medium term contract vacancies that would help with some of these issues.

2 OARC Services Overview

2.1 Data Archiving

OARC maintains a large store of multiple data sets.

Day in the Life OARC coordinates annual and occasional ad-hoc Day in the Life (DITL) DNS traffic capture events. These involve many operators of significant DNS infrastructures—including root server operators, TLDs, and recursive operators—running packet captures of their traffic over the same 48 hour period. The data are uploaded to OARC where it is organized for use in research.

The DITL collections go back to 2009.

RSSAC 002 Statistics The Root Server System Advisory Committee's publication RSSAC 002 is the Advisory on Measurement of the Root Server System. It defines an initial set of statistics to be collected by root server operators from their systems. OARC collects the output of this reporting from each root server operator, daily, and maintains a history of these statistics available for analysis or review.

Zone File Repository OARC maintains an historical archive of zone files which includes daily updates of the root zone going back to 1993, and weekly updates of several TLDs beginning at various times between 2009 and 2018.

Other Data OARC also periodically accepts submissions of other data that may be relevant to researchers interested in the DNS:

- derivative data from research done on OARC's other datasets
- data collected from OARC testing tools, such as the DNS Entropy Tester
- DITL-like collections from outside regular DITL windows, such as occasional contributions from AS112 server operators
- packet captures from OARC's Open DNSSEC Validating Resolver (ODVR) which includes forwarded queries from the DNS Privacy Testbed

- Case Western Reserve University's "Case Connection Zone" FTTH data
- other ad-hoc contributions of relevant data

2.2 File Servers and Storage

OARC's datasets are stored on six file servers. The first five file servers, located in Fremont, California, have 424.31TB used of their 532.64TB of capacity. Two of these have multiple filesystems, marked as A and B in the chart below. The sixth file server, located in Ottawa, Ontario, is an off-site copy of a selection of datasets from the first five servers.

Server/Volume	Used	Capacity	Notes
FS1	118TB	121TB	Offline - RAID failure
FS2a	36TB	42TB	
FS2b	40TB	125TB	
FS3	34TB	42TB	
FS4	72TB	84TB	
FS5a	69TB	84TB	Offline
FS5b	33TB	42TB	Offline
FS6	117TB	121TB	Located in Ottawa, Canada

Each file server uses either ZFS (RaidZ2) or XFS over software RAID for its filesystem to provide redundancy within the file server. Each dataset is stored on more than one file server in order to create cross-chassis redundancy of data; some datasets currently have copies on three systems. This means that the total size of all unique datasets is slightly less than half of the 504TB indicated above.

All capacity numbers above are the filesystem capacity, rather than the raw size of the disks in service.

The above data does not include the 2021 DITL collection, which was recently completed, and currently resides on temporary storage before being moved to `FS2b`. The un-processed RAW data is roughly 13TB in size, and its post-processed copy is expected to be between 6TB and 11TB.

These servers are all due to be replaced in the coming months with the new Ceph storage cluster, discussed in detail in previous reports, and covered again below in section 3.8.

2.3 Data Analysis Servers

OARC maintains four UNIX shell servers with access to the above data sets. Three in Fremont, CA (an1, an2, an4) and one in Ottawa, ON (an3). Members and Supporters who have signed a Data Sharing Agreement and request access are given accounts on these analysis servers, which they can use to do research into the DNS using any of OARC's datasets.

Note Well: No data, even derived data, may leave OARC analysis systems without express written authorization, in compliance with the Data Sharing Agreement. Contact admin@dns-oarc.net first, *always*.

3 System and Service Status

3.1 General Condition

I'm happy to report that the general stability of OARC's network and systems seems to have reached a stable state. There remain a number of systems and services that require more care and attention than I would like, but it seems that the days of major, disruptive surprises are past us. The last six months has seen a number of high-maintenance systems and services either cleaned up or replaced. We have more to do in this regard, but the list is steadily shrinking.

Much of our regular hands-on work continues to be disrupted by the pandemic, but it seems than an end is at least in sight, and by Q4 at the latest we should be back to normal with respect to on-site visits and physical maintenance.

Although we have not yet posted a new RFQ, we continue to be on the lookout for a long term contract with someone who is able to act as emergency remote hands, however with travel restrictions coming to an end soon, the work required for such a contract will truly be emergency response, and not handling any of the more regular physical maintenance tasks we were envisioning late last year.

We are also continuing to search for someone to handle some short term contract work related to routing configuration cleanup, and replacement of our border network gear. Please see the OARC careers page for more detail.

3.2 Authoritative DNS Improvements

OARC has been running the same set of hand-crafted shell scripts for zone signing since approximately 2013. This predates useful auto signing and key management in most (all?) authoritative DNS software, and is likely the reason that no keys have been rolled in all that time. Our internal DNS setup has been, up to now, the very essence of the the English adage about the cobbler's children having no shoes.

We now have a new hidden authoritative server which is making use of the tools in current releases, allowing us to improve our signature and key management, begin rolling keys on a regular basis and, at least as important, begin to roll algorithms.

The first zone imported into this hidden primary was `dnsviz.net`, for which authority has now been migrated from Sandia National Labs to DNS-OARC.

Moving the zone will allow us more flexibility in maintenance for the DNSViz service in the future.

As other zones are migrated from the old primary to the new, we will begin doing regular key rolls, and migrate from RSASHA256 to ECDSAP256SHA256 for all zones. Once the authoritative data of all of our zones are migrated to the new server, the old `ns1.dns-oarc.net` will be decommissioned, and the new currently-hidden primary will become public.

3.3 General Cleanup

Several more applications have seen cleanup, documentation, and packaging as necessary to allow me to migrate them from older servers. The process of emptying out older systems so that they can be reinstalled or recycled continues, and we're getting closer to the goal of eliminating the older, poorly documented services and processes.

In the past two quarters we've also seen major OS updates to a few systems, reducing the number of Devuan 1.x hosts, and eliminating more hand-compiled pieces of software that should be installed from packages.

3.4 Fully Automated (OS) Installation

Over the winter I have been working to set up FAI (Fully Automatic Installation), software designed to automate OS installs. It's somewhat reminiscent of Kickstart/Jumpstart but with a bit more of a rules-based approach. It uses a site-defined list of classes, each of which comprise certain rules for system setup, which can be combined in different ways to create an install rule-set for any individual system.

The fresh urgency of the analysis systems upgrades became our real trial of the system, and it worked well.

Having our OS install rules baked into the configuration of FAI, as opposed to being a documented procedure to follow, will make server setup much more deterministic and repeatable. Use of FAI also speeds up the process of server reinstalls significantly. The combination of these two improvements, along with future use of more configuration management tools, will significantly reduce the ops impact of server replacements, additions, etc.

3.5 Analysis Servers

The analysis servers have been due for an OS refresh for quite some time. We've been putting it off because of several combined factors: the servers are in near constant use, making scheduling the down-time complicated; the hardware is extremely old, and our usual tools for a remote OS install won't work; the servers are extremely old, and rely on firmware drivers no longer in the generic install kernels of current OS releases; physical visits are currently impossible.

The analysis servers were among the last remaining Devuan 1.x systems on the network, and have not even been seeing security patches for the last year and a bit. And so, when the Baron Samedit `sudo` exploit came along in February, that raised the priority of an OS reinstall high enough that that we had to find an immediate solution.

As discussed above, I had already been experimenting with FAI, and ended up using that to solve the problem for us.

The remaining server in Ottawa will see an OS update soon, once I have time to clone our FAI setup in that site.

> **Important Note**: Now that the analysis servers are patchable, we are setting a consistent maintenance window on the 1st and 3rd Tuesday of every month, between 15:00 UTC and 19:00 UTC. During this window we will apply any waiting patches, and *may* reboot the servers if necessary. Reboots will not be announced in advance, however they will be set on a 30 minute timer with a `wall(1)` notice going out to all logged-in users at the start.
>
> In the event of critical patches that cannot wait for a regular window, we will send a notification to the members' mailing list in advance of maintenance.
>
> In the past, researchers have been able to leave analysis scripts running indefinitely, sometimes for weeks at a time. These may now be interrupted due to reboots, and so researchers should take care to make their data analysis restartable.

3.6 Discontinuing the Shared DSC Platform

It's now been nearly two years since DNS-OARC replaced our member portal, and as a consequence removed the only way to view the shared DSC data that OARC has been collecting for many years. In that time, only one member—and DNS-OARC itself—have been contributing data to the shared DSC platform. I have made several requests for feedback on the shared DSC data at workshops over the past few years, and have yet to receive any comments.

As a result, I have recommended that we discontinue collecting data for a community DSC instance. Once the final decision is made, we will be in touch with the member still contributing data about shutting that down.

3.7 Networks, Routing, and Routers

The routers OARC uses in Fremont and Ottawa are now beyond end-of-support. We were eventually able to obtain the final IOS update for them in Q1, and complete a router update. Unfortunately, the patch did not solve our ongoing v6 routing issue, so we continue to have to frequently reset our main transit

session, and occasionally reboot the router, in order to maintain reachability to the entire v6 Internet.

We are still planning to replace our routers in the first half of the year. Due to engineering workload, we're hoping to offload the task of hardware selection and configuration to a short term contract. Details are available on OARC's web site on our careers page.

Once we have new hardware in place, we have plans to improve OARC's network even further by setting up our first redundant transit session. We have received an offer from Mythic Beasts, who host OARC's Mattermost server, to take advantage of a new transit service they are providing in our existing data centre.

3.8 File Servers

The Old Stuff

As mentioned in the previous few Systems Engineering reports, OARC's file server infrastructure is aging and beginning to experience an increasing frequency of hardware related issues. These include, but are not limited to, crashes and data errors when the systems are put under load.

In the last six months we have stabilized `fs2`, but have lost the RAID on `fs1` (from catastrophic failure of the disk controllers, resulting in corruption of the RAID volume), and have seen `fs5` begin to have problems booting.

We had already made the decision in the second half of 2020 to reduce the amount of time and budget that we were investing in maintaining those systems. Our focus has instead been on reducing the load on the old servers as much as possible, in order to minimize the likelihood of any further failures, and to try to focus on the replacement infrastructure.

Our data duplication procedures, designed to ensure that the loss of any individual file server does not result in data loss, have so far protected us from any actual loss of data. Keeping a minimum of two copies of every data set means that, even with the failure of `fs1`, each of its datasets are still available on at least one other server. While `fs5` does not currently boot, its storage volume was in fine condition when it was last up. When it comes time to move data off that server, we expect to be able to cannibalize hardware from other systems in order to make it bootable, or transplant its drives to another system. In addition, many of the datasets from `fs1` have a third copy on `fs6` in our Ottawa satellite location.

The problem of off-site replication of the data housed on our file servers remains an issue. We are constrained by both the Data Sharing Agreement and by available budget. The new storage infrastructure increases our on-site replication from 2x to 3x for each block of storage, and we are budgeting for more regular hardware refreshes of the new infrastructure than the old one ever saw, both of which should protect us against the kinds of problems we're currently having with the old infrastructure. However, catastrophic failures

still happen, and it's important that we find a solution to keeping some sort of off-site backup of these important datasets.

The New Stuff

For over a year now, we have been working on the design, procurement, and setup of a new Ceph clustered storage system to place the existing file servers. We've seen several delays in that time, for various reasons, but continue to make steady (if slow) progress.

Some of our hardware shipments in the fall were delayed until just before the winter holidays. Despite that, we managed to get the bulk of the hardware racked and cabled before the end of the year. At that point we discovered that the power budget we'd received from Dell for the servers seemed to be the idle draw, not maximum draw, as we'd been told. After some troubleshooting, we confirmed that was the case, and so will be required to temporarily double the number of circuits we have, and do some re-cabling. Once easy travel is again possible, I'll use the new information we have about power draw to reorganize the servers between cabinets, allowing us to spend budget on power more efficiently.

Our fibre switches for the 10G cluster network were the last item to arrive, and have now been racked, but not yet cabled or configured. We are now resigned to using the more expensive data centre staff to get that work completed, and will be making plans to get it done during Q2.

We're still anticipating having the cluster up and running, and beginning to migrate data to it, by the end of Q2.

3.9 Day in the Life Dataset

Readers may recall that in the fall we were just beginning to run the post-processing on the most recent three DITL collections, having finally stabilized our only file server with enough storage to hold them.

The DITL datasets that we collect go through a post-processing step before being made available to researchers. The process standardizes the PCAP files around a few simple assumptions:

- consistent layer 2 information, to simplify packet processing
- consistent start and stop times, at 5 minute intervals, for every PCAP file
- consistent compression algorithm and settings
- time-ordered packets in each PCAP file (PCAP captures from the network do not guarantee packet write order)
- root operator data stored according to the root server name rather than the operator name

- de-anonymization of root server addresses for those root operators that do query anonymization affecting the server as well as client address

This processing takes two to three weeks per DITL collection, and puts tremendous read/write load on the file server while it is taking place.

The post-processing for the 2018 Root KSK Roll, and regular 2019 and 2020 DITL collections were all completed late last year, after being delayed by stability issues with the storage hardware. Recently some questions have been raised by one researcher about apparent low volumes in the cleaned vs. raw data for one root operator's contribution, which may indicate a problem in the way that post processing is being done. I'm continuing to investigate.

There have been more questions this year about meta-data for DITL contributions, and we are thankful to Mark Allman for presenting on meta-data that DITL contributors can include with their data to aid researchers. We don't have a formal process for receiving such meta-data, but have been including a README with each contributor's data, if they supply it, for many years. DITL contributors can provide meta-data about their contribution by emailing the relevant information to admin@dns-oarc.net, or to the DITL mailing list.

The 2021 DITL collection was run three weeks ago, and has been an excellent success. We were receiving long-tail uploads from contributors as recently as the last few days. So far we have received approximately 13TB from 22 contributors (as compared to 16TB from 21 contributors for 2020). Once we confirm that all contributors have completed their uploads, we'll start the slow process of doing post-processing on this newest collection, which we expect to be available to researchers after slightly more than a month.