Systems Engineering Update
OARC AGM 2020, The Ether

Matthew Pounsett

2020/09/22

Contents

1 Introduction

The past year has seen a number of challenges, both technical and otherwise, that we've had to surmount in order to make progress on the engineering side of things. Although we have a shorter than usual list of wins to lay claim to since OARC 31, we have made a surprising amount of progress on some things, and have managed to hold a significant amount of chaos at bay with others.

Details are covered below, but in the past year we have, among other things:

- designed and ordered the equipment for an entirely new, modern storage infrastructure;
- stabilized the existing storage equipment enough to concentrate on other issues;
- enabled full features on DNSViz and replaced its back-end database.

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 24 hour period. The data are uploaded to OARC where it is organized for use in research.

The DITL collections go back to 2009.

DNS Statistics Collector DSC is a data collection and statistics generation tool for DNS. Several members contribute their DSC data to OARC, and we make this available to all members to view on our centralized DSC installation.

There is more information on DSC at `https://www.dns-oarc.net/oarc/data/dsc`.

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.

File Server	Used	Capacity
FS1	118TB	121TB
FS2a	36TB	42TB
FS2b	40TB	125TB
FS3	34TB	42TB
FS4	72TB	84TB
FS5a	69TB	84TB
FS5b	33TB	42TB
FS6	117TB	121TB

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.

At the moment, several file servers are either offline or have been disconnected from the analysis systems due to hardware issues. More details on this are available below in section 3.6.

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

Except for a few notable exceptions, OARC's systems and network continue to improve in stability and reliability. Unfortunately, we continue to experience several disruptive issues which impact our ability to spend time on more strategic improvements to OARC's systems and network.

Some of the more visible and time-consuming work is discussed below, as are several improvements made in the last year.

Unsurprisingly, much of our regular work has been disrupted by the pandemic currently affecting everyone. As I am currently unable to travel to the US, my regular visits to our Fremont site have not been happening. During the early part of this year we had a temporary contractor available to deal with some of our hands-on work there, primarily in the area of maintaining our failing file server infrastructure, but that contract has ended.

We are currently searching for someone to take on a regular contract to handle the sort of physical maintenance that is either too expensive or too complex to delegate to the data centre NOC. In the past we have relied on friends of OARC for volunteer work when there are emergencies, but pandemic travel restrictions mean that the current work list goes well beyond emergency maintenance. We are uncomfortable asking volunteers for the amount of on-site support currently required, and are spending more than we would like on data centre remote hands for routine physical maintenance. If members know of individuals physically located in the Fremont, California area, that regularly do contract network and server operations work, we would appreciate

introductions being made. For more specifics about our requirements, please speak to Matt Pounsett or Keith Mitchell.

3.2 Secondary DNS Migration

It was announced just prior to OARC 31 that we were in search of a new secondary DNS service provider, as our previous provider (ISC) was shutting down their shared platform after many years of pro-bono service. Several members provided us with effectively identical quotes, which forced us to make a very difficult choice. After some deliberation we settled on the UltraDNS service at Neustar. The migration went smoothly, and we have been quite happy since it was completed.

3.3 Shared DSC Platform

The new implementation of the member portal, launched last summer, does not include the joint presenter for the DNS Statistics Collector (DSC) data that the old portal had. In recent years the number of contributors has declined to just RIPE NCC and OARC itself, and the number of people viewing this data has declined even more.

We are considering options for the future of this data set. As discussed in the last two Systems Engineering reports, one possible option is to stop accepting DSC data from members entirely, and archive the existing dataset. Another option is to continue to accept data, and set up a new presenter, possibly based on Grafana, as future work. We have not heard from any members since OARC 30 (May 2019) about their interest in this service; we are still interested in feedback from members about how useful they find this service, and what path we should take with it in the future.

In the absence of any guidance from members we are likely to make a decision within the next six months based entirely on staff workload.

3.4 Networks, Routing, and Routers

The routers in use at OARC's two sites are quickly approaching their end of life. The ASR-1000's we're using are limited to a particular branch of IOS which we understand will see its last updates in the coming months.

At the same time, the presence of DNSViz has increased the visibility of IPv6 issues at our California site, and has made it apparent that we're having intermittent issues with IPv6 routes not being installed properly in the FIB. A root cause and solution for this problem has not been confirmed, but it's been suggested it may be related to a known problem with the version of IOS we're currently using.

OARC's routers are not under any kind of support contract, which further complicates the issues we have. We currently have no way to obtain the minor IOS updates that are available for the ASR-1000 platform.

We're intending to solve all of these problems by replacing our routing hardware in early 2021. Budget limitations mean we will likely not be purchasing new hardware, and what we do get will likely be in its final years of support, but we're aiming for hardware that is still supportable, and comes with access to OS updates, which means we should still see an improvement in the supportability of our networks.

We have also been in discussion with Mythic Beasts, a hosting provider based in the UK who have recently deployed new infrastructure in the same data centre we use in California, to add additional transit supplied by them. Initially this will only be for backup administrative access, and at a lower bandwidth than our main transit link, but we are hoping for the opportunity to promote it to a fully redundant transit service once we have new hardware in place that can support the additional BGP routes.

3.5 SSL-based Services

In the past, OARC has depended primarily on a single wildcard SSL certificate for securing most of its web sites, with only a couple of individually named certificates to cover sites that didn't fit the wildcard. This ended up resulting a few awkwardly named services, and in one that did not initially have an HTTPS version.

During the past summer we've transitioned entirely to using LetsEncrypt for securing OARC's web sites, and mail services, increasing our flexibility, and eliminating several expensive commercial certificates.

3.6 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.

During the past year, OARC has invested a large amount of time internally, and also engaged a temporary consultant for several months, with a view to limiting the impact on staff Engineering time. Those time and financial investments have paid off in some places, and not in others.

While past issues with the ZFS volumes on `fs2` and `fs5` have been mitigated, we have had poor luck in some other areas. Despite a significant amount of work going into the machine to stabilize its volume, it appears that `fs1` may have finally had a complete failure of its RAID (this is a recent event and is still under investigation to confirm whether it is recoverable). And while `fs5`'s volumes were eventually stabilized, the server itself is not currently booting due to a different hardware problem (a failed motherboard or processor). `fs5` is known to have been dropped while being donated to OARC; obvious physical

damage to the server has been a long-standing suspected cause of several odd behaviours it has exhibited.

Our data duplication procedures, designed to ensure that the loss of any individual file server does not result in data loss, should protect us from the current set of hardware issues. Keeping a minimum of two copies of every data set means that, even if the failure of `fs1` is permanent, each of its datasets should be 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.

We have made the difficult decision to invest as little new time, effort, and budget as possible in the old infrastructure. As discussed at OARC 31 last fall, we have been pursuing the deployment of a more modern storage infrastructure, which we expect to be in production around the end of the year. Investing more time than necessary in the old infrastructure would only serve to delay the migration. As a consequence, some datasets will remain unavailable until the new storage infrastructure is in place.

The problem of off-site replication of the data housed on our file servers is still 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

Last year, a major topic of discussion was the need to design, purchase, and deploy a new storage infrastructure for OARC's nearly 300TB of data. By the end of the first quarter of 2020 we had arrived at a final design, and settled on specifics for the hardware deployment, but had reached an impasse with our vendor of choice. Dell had provided us with highly attractive quotes for the hardware we required, but had insisted on commercial terms that were unreasonable for the sum of money changing hands.

Over the past six months we've had a difficult time finding a way around that impasse, but with the help of some members we have now found a reseller who was not only able to improve on Dell's pricing, but also provide us with reasonable terms of sale. We're happy to report that the server hardware is now nearly all on site (the remaining servers should have arrived by the time OARC 33 begins) and that the other hardware bits required (switches, cabling, etc.) are all on the way.

We are still in search of a contractor capable of handling the initial racking, cabling, and base configuration necessary for remote access so that I can begin

the process of installing and configuring the new storage infrastructure. We have a number of leads which look promising, but if Members are aware of network and systems operations contractors in the Bay Area we would appreciate an introduction.

Provided that we can get a base configuration for remote access in place soon, we anticipate that the storage infrastructure could be configured, tuned, and monitored sufficiently for production use by the end of the year.

3.7 Day in the Life Dataset

Owing to the hardware problems we've been experiencing with the storage infrastructure, we are nearly two years behind on making our new DITL datasets available to researchers. The only file server with sufficient free storage to hold the new datasets is `fs2`[1], and until recently it had been experiencing significant write errors on any attempt to process large volumes of data.

Now that the server's storage volumes have been stabilized, we have finally been able to complete the regular post-processing work for the October 2018 collection made during the roll of the root zone KSK. The DITL datasets that we collect go through a processing step before making them available to researchers that standardize 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
- 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 2018 Root KSK Roll collection's post-processing is complete as of Monday, September 21st. We anticipate that the regular spring 2019 DITL collection's post-processing will be done in 2-4 weeks, and that the 2020 DITL collection will be available to researchers in 3-6 weeks (by early November).

The search required by the legal case referred to in the President's report will consume some of our processing capability. This may delay some post-processing of the two more recent datasets.

[1] Yes, this means that new datasets received since the 2019 DITL do not have cross-chassis backups.

3.8 Analysis Servers

Our analysis systems, open to any Member or Supporter for research, continue to see heavy use. We are often put in the position of needing to remind researchers that our analysis systems have constrained resources that must be shared. As suggested in the previous report, budget that was put aside for expanding storage on the analysis systems has been put toward improving the network storage infrastructure, but we intend to return to this issue in 2021.

The OS refresh of the analysis systems, which has been put off due to time constraints, is becoming more of a problem, however, and will need to be addressed within the next couple of quarters.

3.9 DNSViz

During OARC 31 we were in the process of completing what we believed were the final steps in recovering the old DNSViz database, but actually turned out to be only the midpoint in a long series of difficult problems caused by the combination of trying to process an extremely large volume of data in very constrained storage.

Since then we have deployed a new database server, which is running the DNSViz service without the old historical data (but now operating with full features, collecting new history), and made the decision to back-burner the recovery of the old data.

Our original strategy for data recovery was to focus on solutions that could get us results the fastest, which turned out to be the wrong approach entirely. We have a new strategy in mind which, while extremely slow, should be relatively risk-free. Other, far more urgent issues have monopolized our time in recent months, hence the decision to back-burner the recovery of the old historical database. However, we plan to resume those efforts in the new year.

During the past year we've also implemented a workaround for the long-standing IPv6 peering dispute between Hurricane Electric (our only transit provider) and Cogent Communications, the operator of C-root. The lack of a v6 path between HE and Cogent has been the cause of many problem reports with DNSViz, as it caused end users receive errors related to the reachability of C-root, and occasionally their own name server infrastructure. Earlier this year we worked with Cogent's network operations to set up a tunnel between OARC's California site and one of the nearby C-root instances. This does not solve general reachability problems between HE and Cogent, and only makes C-root itself visible to DNSViz, but significantly reduces the number of problem reports that the peering dispute generates for OARC.

3.10 Zone File Repository

In early 2019 ICANN moved its Centralized Zone Data Service away from having zones accessible at static URLs to a RESTful service. OARC's Zone

File Repository has always been operated by some fairly simple shell scripts which were not capable of this more complex processing.

This change was able to pass OARC by without anyone noticing. We suspect this is because any notice of the change likely went to the defunct email address of a previous admin.

This introduced a gap in OARC's zone repository which we did not notice until early in 2020. In February this year we began a project to re-implement ZFR to support static URLs, TSIG-secured AXFR, as well as ICANN's REST service. While the code was finally completed late this summer, we have not yet been able to get it into production due to time constraints. We're hoping to complete this step in the next month or two, time permitting.

In addition to supporting several zone download methods, the new code also supports pluggable filters (necessary for some zones which are consistently malformed), and support for monitoring of download status, so that we can be made aware of zone update failures in the future. Look for a blog post in the coming months describing the new archive capabilities.

Given the recent push to move many older zones over to CZDS, we are now in the unfortunate position that our archive of several zones contains growing gaps in data. We will be reaching out to the community for any contributions of data which would be able to reduce the size of these holes after the new code is in place.

3.11 Open DNSSEC Validating Resolver

For a little over ten years, OARC has been operating the Open DNSSEC Validating Resolver service (ODVR), which consists of one instance each of BIND and Unbound, open to the Internet. The service was originally intended as a way to allow people to test DNSSEC operations and software implementations in a time when validating resolvers—in particular open validating resolvers—were few and hard to find. In 2020, public open resolver services almost universally do DNSSEC validation, and it is much easier (and often the default) for any newly configured resolver to do validation.

Given the overhead necessary to run an open resolver safely (ensuring that it does not get abused, and dealing with attempts to abuse it), and comparing that with the easy availability of validating resolvers, OARC has been considering whether it is worth continuing to provide this service. Last month, we out a request to the entire DNS community for anyone relying on the service to contact us so that we could discuss how to continue to provide the service in a way that is both useful and less of a drain on OARC's engineering resources.

We have not been contacted by anyone in the community.

Based on the lack of response, we have determined that the best course of action is to discontinue the ODVR service, which will be shut down at the end of September, 2020.

This will also have a small impact on OARC's DNS Privacy Testbed, which currently forwards all of its queries through ODVR. The Privacy Testbed will

be reconfigured to remove forwarding from DNS resolution. The DNS Privacy Testbed has not been subject to the same kind of query load as ODVR has recently seen, however its status as an open resolver will be quickly reviewed should we see any issues with it.

4 Conclusion

Some significant challenges have consumed a large amount of my time and attention in recent months. However, I believe we're about to push past those distractions and regain much of the momentum we had last fall and winter. I expect we should have some very visible changes to report by OARC 34, in the spring.