Clustering the OARC Storage Infrastructure

Matthew Pounsett

2019/10/27

Contents

Contents **1**

1 Executive Summary **2**

2 Justification **2**
2.1 Current Architecture & Conditions 2
2.2 Summary of Current Challenges 4
2.3 Advantages of a Clustered Filesystem 5

3 Requirements for a Ceph Storage Architecture **7**
3.1 Total Raw Storage 7
3.2 Server Roles 7
3.3 Object Storage Servers 8
3.4 Meta-data Servers 9
3.5 Monitors 9
3.6 Managers 10
3.7 Data Processing 10

4 Hardware Selection and Budget **11**
4.1 Server Types 11
4.2 Quoted Prices 11
4.3 Platform Selection 11
4.4 Draft Budget 13

1 Executive Summary

DNS-OARC's existing infrastructure for storing its main asset, the collection of *Day in the Life* packet captures and other DNS measurement data, is not only aging in a physical sense, but is also using an outdated architecture which results in a great deal of lost time and extra expense. OARC is proposing that we replace that existing file server infrastructure with a modern clustered filesystem.

Bringing OARC's existing hardware up to date with the required upgrades to CPU and memory, and replacing aging hardware, would cost approximately US$80,000. Instead of spending that capital on bringing older hardware up to date without updating the architecture, OARC is proposing that we instead redesign the infrastructure. Replacing the existing design with a Ceph clustered filesystem is expected to cost up to US$125,000[1].

2 Justification

2.1 Current Architecture & Conditions

OARC's current data storage architecture is based on six separate file servers holding eight discrete volumes of data (two file servers have two volumes each). Each volume is shared with data consumers—such as the analysis shell servers—as an individual NFS[2] volume. Some off-site duplication is achieved by having one of these file servers in a separate location, however the volume on that server is not sufficient to hold all the data from the other seven volumes, and so it is not a complete off-site backup. OARC accomplishes on-site redundancy by using ZFS (RAID-Z2) or software RAID (RAID60) on each volume, and by duplicating each dataset across multiple chassis.

OARC's file server infrastructure holds approximately 230TB of unique data. Manual duplication of these data across chassis brings the current total storage in use up to approximately 510TB. OARC acquires an additional 10-15TB of data with each DITL[3] and DITL-like data collection. On occasion, such as in 2018 and 2019 with the root KSK roll, multiple data collections may be run due to special circumstances, resulting in a doubling of OARC's data increase for the year. The size of a DITL data collection also grows each year as global query rates rise, and as the number of participating operators increases.

[1] This amount is arrived at using Dell US's list price for hardware, minus 37%, which is the typical discount on Dell hardware that OARC enjoys. Dell's prices are used for budgeting purposes only, and have been chosen for the ease with which they can be obtained and compared.

[2] Network File System

[3] Day in the Life: an annual event where large authoritative and recursive DNS infrastructure operators collect packet captures from their servers over the same 48 hour period.

All combined, the file servers currently have approximately 1.2 petabytes of raw disk[4]. After accounting for RAID parity, other redundancies, and spares, there is approximately 660TB of available storage.

Extrapolating the 510TB of actual data stored on 660TB of available storage, and comparing that to the amount of actual unique data (230TB) and underlying raw storage (1316TB), it can be calculated that the effective redundancy multiplier of the RAID and ZFS filesystems, combined with cross-chassis duplication, is just under 4.5x (442.13%).

$$\frac{1316\text{TB}(\frac{510\text{TB}}{660\text{TB}})}{230\text{TB}} = 442.13\%$$

At present, OARC is using 12TB drives in its annual volume updates, which, in 2018, cost approximately US$10,000 per 25 drives (for the 24 drive file servers), or US$18,400 per 46 drives (for the 45 drive file servers). At current prices, these drives are available at around US$8,000 per 25, or US$14,700 per 46. Drives are currently forced to remain in service for up to eight years before regular replacement.

Although OARC regularly replaces the hard drives in the existing file servers, the servers which comprise the existing infrastructure are all overdue for upgrade or replacement of other components, or the whole chassis. Several of these file servers could get by with only component upgrades (CPU, memory, or both) but are old enough that we would have to pay a significant premium on acquiring old components to do so.

One server suffered severe physical damage in transport when it was originally being donated to OARC. In addition to making it difficult to swap drives in this system, the damage makes it dangerously insecure in its cabinet, and seems to be at the root of some strange behaviours of the system.

By replacing all six file server chassis, and upgrading the drives in five of them to new 12TB disks, we could update our existing equipment for approximately US$80,000[5]. Upgrading all of the drives compensates for decreasing the number of drives in the two 45-drive chassis, and for retiring the two external JBOD chassis. This assumes that the Ottawa file server is retired, eliminating what little off-site backup we currently have, and does not address the cost of future expansion or the cost of any of the operational challenges with the current architecture.

OARC is proposing to replace that infrastructure with a modern clustered filesystem, comprised of a larger number of smaller servers than the current architecture calls for. The initial capital expense of the transition would be higher than a simple modernization of the current architecture, but solves many of our operational challenges, and could reduce (or at least smooth out) the future cost of expansion.

[4] 161 drives of 4TB, 48 drives of 8TB, and 24 drives of 12TB, totaling 1316TB of raw disk.

[5] This is the approximate cost of six iX-4224 servers from iX Systems, with 512GB of RAM, plus sufficient drives to replace all those disks smaller than 12TB and fill all six servers (120 disks).

2.2 Summary of Current Challenges

Reliance on Vertical Scaling

In order to keep growing this infrastructure, OARC must annually upgrade all disks in at least one volume, using whatever commodity disk drives provide the best cost per byte at the time of the upgrade. This sort of upgrade path occasionally requires an increase in CPU or available RAM, especially on the ZFS-based systems where RAM needs to track available storage in order to provide efficient write caching. Over time this can cause the server to require memory sticks or CPU chips that are no longer "commodity" components, and have very high cost per megabyte (for memory) or performance cost (for CPUs).

OARC is currently many years behind on tracking CPU and RAM upgrades in the file servers. At this time, several servers need complete replacement, and while others could potentially manage with only a CPU and/or RAM upgrade, their motherboards are now so old that purchasing the necessary components is subject to price premiums due to limited supply.

It is expected that the most economical way to bring the file servers up to date will be a complete replacement of all six chassis with current hardware.

OARC's acquisition of new data appears to be slightly out of pace with the equivalent to Moore's Law for storage technology. Occasionally, OARC is forced to add an extra volume to the file server infrastructure instead of doing a drive upgrade on an existing volume (this is how we arrived at having two more volumes than we have file servers). This can happen either because a simple upgrade would not provide enough additional storage for the coming year, or because there is insufficient free storage in the other file servers to vacate the volume which needs upgrading.

Adding a JBOD drive shelf to an existing server is a cheaper option since a disk chassis is about half the cost of a new server. However, this has the effect of putting more load on the server it is attached to, driving its memory and CPU requirements even higher. Furthermore, adding a new chassis (whether a JBOD drive shelf or a new server) has the additional effect of either extending the number of years disks must remain in service[6] or causing large step functions in OARC's capital budget in order to maintain a fixed service lifetime for disk drives[7].

Inefficient Use of Storage

Individual datasets are quite large, on the order of tens of terabytes each. A "best fit" arrangement, packing servers as tightly as possible with data, still

[6] As OARC currently only upgrades one volume per year, drives must remain in service for up to eight years, unless replaced due to failure.

[7] Assuming 12TB drives at the current price for a full cycle of replacement, and a five year maximum lifetime, that is US$15,000 for each of the first two years, and US$16,000 for each of the remaining three years given the current number of volumes. As the number of volumes increase, individual years' replacement costs jump by US$8k or US$15k depending on the size of the new volume.

leaves a large amount of storage unusable. At present, there are about 50TB of available storage across all of OARC's file servers which cannot be filled with any existing dataset.

High Manual Labour Cost

There is a high cost in time associated with manually duplicating datasets across chassis, calculating "best fit" to use storage as efficiently as possible, and periodically verifying that datasets remain true duplicates. Individual datasets are quite large, and can take days to copy from chassis to chassis even across a 10Gb network. Generating checksums of files, even with a relatively low-complexity algorithm like MD5, can take weeks per file server.

Some datasets, such as the Zone File Repository, continually grow. As they increase in size they can outgrow the volume they reside on, and need to be relocated. In addition to the labour cost of moving the data, this often results in the need for carefully planned and carefully timed reconfiguration of software in various places that either write or read the affected data.

As any volume has its drives updated, it is necessary to move the data off of that volume onto others so that the drives can be swapped and a new, larger filesystem created in its place. This relatively frequent re-balancing of data causes things like best fit calculations to have to be repeated on a regular basis.

Complexity for Researchers

Since datasets are manually duplicated across chassis, there is more than one path to each dataset, which can be confusing to anyone approaching OARC's data for the first time.

All of these operations that involve the movement of datasets between volumes make the path to datasets on the analysis systems a moving target for researchers, which can cause existing research tools to fail in unexpected ways.

2.3 Advantages of a Clustered Filesystem

Transition to Horizontal Scaling

Rather than relying on a small number of very large file servers, clustered filesystems tend to be designed around a larger number of much smaller systems. This has several advantages.

- Servers always stay firmly within the zone of "commodity" hardware. They are never required to support and power large numbers of drives, eliminating the need for expensive chassis, drive controllers, and high wattage power supplies, and the modest performance requirements of each individual server allow memory and CPU can be selected from the most common (and therefore cheapest) components.

- As the failure of any individual server in this situation is less of an impact, many of the redundancies required by monolithic file servers become unnecessary (e.g. dual power), further reducing the capital cost.
- Expansion of the cluster is no longer a huge step function. Since the servers are smaller, with less storage per server, expansion can be done in much smaller, less expensive increments.
- Since clustered storage does not rely on traditional redundant filesystems (e.g. ZFS, RAID) the requirement to maintain near-identical drives is removed, allowing the purchase of precisely the drives necessary in any year for upgrade or replacement. This allows for a smoother, more predictable growth curve in the capital budget for storage.

Performance

A clustered filesystem distributes reads and writes across multiple chassis, much the same way a mirrored filesystem distributes reads and writes across multiple disks. In addition, Ceph separates block storage from filesystem meta-data, which significantly improves the performance of both seek operations (e.g. `ls`, `find`, `stat` operations) and data reads/writes by putting them on separate disks (and even separate servers) which prevents the two types of operations from impacting the performance of the other.

The distribution of these activities across multiple servers can significantly improve the performance of the filesystem, which will directly result in OARC's researchers being able to get their answers more quickly.

Efficient Use of Storage

Cross-chassis replication and filesystem redundancy are one in the same in clustered storage. The Ceph storage system we are proposing recommends a 3x redundancy, meaning that each block of data would be stored on three separate servers. This is an improvement in the efficiency of the storage from the current state, reducing the amount of raw storage required to support OARC's data.

In a clustered storage architecture, we would transition from eight distinct NFS shares to a single network filesystem. Since there is no longer a need to calculate best fit for datasets on individual volumes, the issue of having wasted storage on volumes is eliminated, further improving the efficient use of the underlying raw storage.

Reduced Manual Labour

Cross-chassis replication in clustered storage is handled automatically as one of the primary features of the software. This eliminates the need for time-intensive duplication and verification of datasets and, combined with the single

filesystem, eliminates the need to calculate best fit for datasets on individual systems.

Simplified Presentation of Datasets

Moving from several separate NFS volumes to a single networked filesystem means that the path to datasets will no longer appear to arbitrarily change as datasets move from volume to volume. Also, since cross-chassis duplication of datasets is handled by the underlying clustered storage architecture, there will be only one path to any given dataset, rather than multiple paths to each copy, which will reduce confusion.

3 Requirements for a Ceph Storage Architecture

3.1 Total Raw Storage

In order to do a direct comparison with upgrading the existing architecture, we will state a requirement for only enough storage to support the current data plus expected 2020 increases. This means we should plan for an initial capacity of 260TB of data.

Automatic duplication in the Ceph storage architecture is recommended to be at least 3x due to the need to retain redundancy even during hardware failure. This puts the minimum raw storage required at 780TB.

The Ceph storage platform will re-balance storage in the event of the failure of a disk or server, attempting to return to 3x replication of all data when access to a copy is lost. This re-balancing requires sufficient unused storage on the remaining systems to do this re-balancing work. We anticipate being able to respond to any hardware failure quickly, however out of an abundance of caution we will plan for the simultaneous failure of up to two systems. This raises the minimum raw storage requirement by whatever amount of storage we have per system, times two.

If we assume the initial deployment should cover three years of growth before requiring expansion, and if we plan for maximum potential growth (i.e. assuming an average of two packet capture exercise per year, and for those captures to be 15TB each) then we should plan for the capability to hold 320TB of data. This would require 960TB of raw capacity, plus two servers worth of extra capacity, using the same calculations as above.

3.2 Server Roles

The Ceph storage architecture has three distinct roles for servers:

- Storage servers, running the Object Storage Daemon (OSD). In this document we will refer to storage servers themselves as OSDs, as shorthand. OSDs store data, handle data replication, recovery, and re-balancing, and provide some monitoring information to Ceph Monitors.

- Meta-data servers (MDSs) store meta data on behalf of the Ceph filesystem. This includes things like the mapping between filesystem path and data objects stored.
- Monitors, or MONs, which maintain the mappings of cluster state, and are responsible for managing authentication between clients and servers.
- Managers, which are responsible for keeping track of runtime metrics and other state information, including storage utilization, current performance metrics, and system load.

As the Monitors and OSDs can both experience fairly intense load, it is strongly recommended that those roles not coexist with any other role on a server.

All servers will be given at least one 10Gb (SFP+) network interface to connect to OARC's existing dedicated data network. If the cost is not prohibitive, a second 10Gb interface may be used to give the Ceph cluster's inter-process communication a separate network from the client/server connections of the network filesystem.

3.3 Object Storage Servers

Ceph requires a minimum of three OSDs to maintain redundancy and high availability, however at the data volumes we're considering this restriction will not be an issue; we require many more than three servers to satisfy our data storage requirements.

CPU

Ceph OSDs coordinate the autonomic distribution of data among themselves, calculate the data placement algorithm, replicate data, and maintain their own copy of the cluster map. Therefore, OSDs should have a reasonable amount of processing power (e.g., dual core processors).

RAM

OSDs that use the currently recommended storage backend require 3-5 GB of RAM. In addition, OSDs should have 1GB of RAM per 1TB of raw storage in order to have enough memory to do re-balancing operations when the cluster size changes (either during failure or expansion).

Storage

Ceph best practices dictate that we should run operating systems, OSD data, and OSD journals on separate drives. As a result, OSD servers will require two modestly sized SSDs for the OS and journals, plus appropriate HDDs to store the OSD data.

Ceph journals have write-intensive semantics, requiring an SSD with excellent write performance for the journal.

Selection of data storage drives will be driven primarily by cost per gigabyte. At the moment, the best option appears to be 12TB HDDs. SSDs provide a significant improvement in access times, but with a commensurate increase in cost per gigabyte. SSDs also have a much smaller maximum storage capacity than HDDs, which has the effect of reducing the maximum storage per server. As the cost of nearly a petabyte of solid state storage is prohibitive, we will restrict ourselves to considering HDDs for data storage.

The number of data drives per server will require a careful balance between maximizing the amount of storage per server (and thus minimizing the number of servers required to attain the required total raw storage), minimizing the cost of RAM (by avoiding costly high-capacity modules), minimizing the cost of the chassis required to support the number of drives per server, and ensuring the total read/write performance of all the data drives is in balance with the total network throughput available.

3.4 Meta-data Servers

Meta-data servers can be deployed singly, but this provides no redundancy or scalability. MDSs can be deployed in active/standby configurations for redundancy and in active/active configurations for scalability. Combinations are also possible, involving more than two servers (e.g. active/active/standby). As our requirements are not expected to be terribly strenuous at first, we will plan to deploy two MDSs in an active/standby configuration.

CPU

Ceph meta data servers dynamically redistribute their load, which is CPU intensive. Meta-data servers are recommended to have significant processing power (e.g., quad core or better CPUs).

RAM

The meta data daemon memory utilization depends on how much memory its cache is configured to consume. 1GB is recommended as a minimum for most systems. This is a fairly modest requirement and will not significantly impact server choice, but must be considered if the Meta-data service will share resources with another role.

3.5 Monitors

Ceph requires at least three monitor servers for redundancy and high availability.

CPU

Monitors simply maintain a master copy of the cluster map, so they are not CPU intensive.

RAM

The Monitor daemon memory usage generally scales with the size of the cluster. For small clusters, 1-2 GB is generally sufficient. For large clusters, we require more (5-10 GB).

3.6 Managers

At least two manager servers are required for high availability.

CPU

The Manager role does not have significant enough CPU requirements to have an impact on hardware selection.

RAM

The Manager daemon memory usage generally scales with the size of the cluster. For small clusters, 1-2 GB is generally sufficient. For large clusters, we require more (5-10 GB).

3.7 Data Processing

Though not part of the list of Ceph roles, we will require data processing servers for this new architecture. OARC currently does data post-processing on DITL and other packet capture datasets after they are uploaded. This work is currently done on each file server where the dataset is stored, so a new home for this role will need to be planned for.

CPU

For this role we will require dual-CPU servers with at least two cores per CPU.

RAM

The RAM required by data processing is not unusually high, but as with all types of data processing, more memory can be helpful with performance. We

will be considering a modest 32GB for systems with this role.

4 Hardware Selection and Budget

4.1 Server Types

After considering the required roles that servers must perform, their relative requirements, and Ceph's best practices for distributing those roles, we have decided on the following distribution of roles into four server types:

OSDs running the object store daemons.

MDSs running the meta data services.

MONs running Ceph's monitoring service.

MGRs performing the Ceph management role as well as OARC's data processing role.

4.2 Quoted Prices

Prices for the above server types are arrived at using Dell US's list price for hardware, minus our standard discount. All prices are pre-tax.

These numbers are used for budgeting purposes only, and Dell's prices been chosen for the ease with which they can be obtained and compared. In order to maintain an apples to apples comparison with a simple update of the existing architecture, this budget only includes enough capacity to support anticipated maximum growth through 2020.

While OARC enjoys a significant initial discount on Dell servers, they still come at a significant price premium. For example, large capacity disks (10-14TB) from Dell are roughly 4x the cost of quality name brand drives from other vendors, even after OARC's discount. For this reason, we are using Dell server prices with OEM disk prices to generate our budget estimate.

We expect that the final cost of this deployment will be lower than the estimates given here. We anticipate being able to negotiate more of a discount from Dell given the size of the proposed order. We also plan to search out multiple quotes before making a final purchase, and expect that a quote from a vendor capable of doing custom builds that more closely match our requirements (e.g. iX Systems, with whom OARC also regularly does business) will be able to provide significant savings over Dell's prices.

4.3 Platform Selection

OSDs

The theoretically cheapest deployment possible for this architecture is, roughly, the one with the most possible disks per chassis in the OSD servers, as the per-

chassis cost hugely outweighs the cost of memory or disk. From a performance perspective, however, OSDs with very few disks per chassis create the best cluster. We need to find the best balance point which gives us a decently performant cluster in an achievable budget.

The ideal performing cluster seems to be around 8 drive slots, with two slots used for the boot and journal SSDs, and six used for storage drives. Building on a Dell R740 platform, with OEM disks, we can build an OSD server with 72TB of raw storage for around US$8,800. However, because of the total amount of raw storage required, this would mean we need to build 13 OSD servers, for a total cost of around USD$115,000 before tax (or US$175,000 using Dell branded disks).

$$\frac{[MinStorage]+2([ServerStorage])}{[ServerStorage]} = [ServerCount]$$
$$\frac{720+2(72)}{72} = 12.8$$
$$13(\$8,800) = \$114,400$$

This is only the cost of the OSD servers, and does not include the cost of the MDSs, MONs, or MGRs. Although the price per server is quite low, and gives us the potential for a very smooth curve in capital cost as the cluster grows, this seems like an unnecessarily high price for the full deployment. As a result, we must instead look at chassis that permit more than 8 drives per server.

Dell's R740xd2 chassis permits up to 24 3.5" drives, which gives us the best possible drive density per server. The price sweet spot for this chassis is at 20 storage drives providing 240TB of raw storage at US$18,500 per server, requiring 6 servers for a full deployment, at a total cost of about US$111,400 for the OSD servers (or US$200,500 using Dell drives). Due to the higher per-chassis cost, this is not a significant improvement over deploying 13 R740 servers. On top of this, deploying servers with so many drives gives us a similar performance penalty as the current deployment, concentrating too much drive streaming capacity behind a single 10Gb/s network interface, and gives us a similarly high step function for capital cost of future expansion.

In between the Dell R740 and the R740xd2 is the R740xd, which can support up to 12 3.5" disks, or 10 storage drives plus the two required SSDs. Even though the R740xd only supports half the drives of the R740xd2, which would increase the number of servers required, it has a significantly lower cost per chassis than the R740xd2. An R740xd with ten 12TB storage drives costs approximately US$10,900, and requires a deployment of ten OSD servers, for a total OSD deployment cost of US$109,000 (US$183,000 using Dell drives).

MDSs

For the Meta-data Server, the best option from Dell appears to be the R540 Intel Silver chassis with an SSD boot drive (smaller and cheaper than the default HDD for the chassis), and an additional 10Gb SFP+ interface for US$2,260 per server.

It would be possible to use a cheaper chassis, such as the R340, except for the need for a 10Gb SFP+ interface, which is not available on any of Dell's 1u Intel servers.

This quoted price does not include the 10Gb/s optics which are not, for some reason, an option to purchase with this chassis. Optics would need to be sourced separately.

MONs

For the Monitor servers, the best option from Dell appears to be the R540 Intel Bronze chassis, with the default disk, a small increase in RAM (up to 16GB), and an additional 10Gb SFP+ interface for US$1,970 per server.

It would be possible to use a cheaper chassis, except for the fact that no 10Gb/s SFP+ interfaces are available on Dell's 1u Intel servers.

This quoted price does not include the 10Gb/s optics which are not, for some reason, an option to purchase with this chassis. Optics would need to be sourced separately.

MGRs

For the MGR servers, the best option from Dell appears to be the R540 Intel Silver chassis, with dual CPUs, increased disk for local scratch storage, an increase in the default RAM up to 32GB, and an additional 10Gb SFP+ interface for US$3,620 per server.

It would be possible to use a cheaper chassis, except for the fact that no 10Gb/s SFP+ interfaces are available on Dell's 1u Intel servers.

This quoted price does not include the 10Gb/s optics which are not, for some reason, an option to purchase with this chassis. Optics would need to be sourced separately.

4.4 Draft Budget

As previously noted, the server prices described here are using Dell servers, based on list prices with OARC's standard discount, with high capacity storage drives obtained from an alternate vendor. Prices do not include tax. It is expected that we will be able to obtain much more favourable pricing through a Dell sales rep, and from quotes obtained from other vendors.

Server Type	Quantity	Unit Price	Subtotal
OSD	10	$10,900	$109,000
MDS	2	$2,260	$4,520
MON	3	$1,970	$5,910
MGR	2	$3,620	$7,240
		Total	$126,670