PROCEEDINGS OF SCIENCE

What Goes Up, Must Go Down: A Case Study From RAL on Shrinking an Existing Storage Service

Rob Appleyard[1]
STFC
STFC Rutherford Appleton Laboratory, Didcot, OX110QX, United Kingdom
E-mail: rob.appleyard@stfc.ac.uk

Dr. George Patargias
STFC
STFC Rutherford Appleton Laboratory, Didcot, OX110QX, United Kingdom
E-mail: george.patargias@stfc.ac.uk

Much attention is paid to the process of how new storage services are deployed into production that the challenges therein. Far less is paid to what happens when a storage service is approaching the end of its useful life. The challenges in rationalising and de-scoping a service that, while relatively old, is still critical to production work for both the UK WLCG Tier 1 and local facilities are not to be underestimated.

RAL has been running a disk and tape storage service based on CASTOR (Cern Advanced STORage) for over 10 years. CASTOR must cope with both the throughput requirements of supplying data to a large batch farm and the data integrity requirements needed by a long-term tape archive. A new storage service, called 'Echo' is now being deployed to replace the disk-only element of CASTOR, but we intend to continue supporting the CASTOR system for tape into the medium term. This, in turn, implies a downsizing and redesign of the CASTOR service in order to improve manageability and cost effectiveness. We will give an outline of both Echo and CASTOR as background.

This paper will discuss the project to downsize CASTOR and improve its manageability when running both at a considerably smaller scale (we intend to go from around 140 storage nodes to around 20), and with a considerably lower amount of available staff effort. This transformation must be achieved while, at the same time, running the service in 24/7 production and supporting the transition to the newer storage element. To achieve this goal, we intend to transition to a virtualised infrastructure to underpin the remaining management nodes and improve resilience by allowing management functions to be performed by many different nodes concurrently ('cattle' as opposed to 'pets'), and also intend to streamline the system by condensing the existing 4 CASTOR 'stagers' (databases that record the state of the disk pools) into a single one that supports all users.

International Symposium on Grids and Clouds (ISGC) 2018 in conjunction with Frontiers in Computational Drug Discovery
16-23 March 2018
Academia Sinica, Taipei, Taiwan

[1]Speaker

PoS(ISGC 2018 & FCDD)026

1. Introduction

The Scientific Computing Department (SCD) at the STFC Rutherford Appleton Laboratory (RAL) hosts the UK's WLCG[1] Tier 1 data centre ('the Tier 1') that forms part of the GridPP[2] project. From 2007 to 2017, the disk and tape storage required for the Tier 1 has been provided solely by CASTOR[3][4], a storage service developed at CERN for managing physics data at multi-petabyte-scale.

In 2017, the Tier 1 brought a new storage system, known as 'Echo'[5] (Erasure-Coded High-throughput Object store) into production. Echo is based on Inktank/Red Hat's Ceph[6] storage system, and is intended to replace some, but not all, of the functions CASTOR performs for the Tier 1.

At RAL, CASTOR provides two distinct kinds of storage service for the Tier 1. The first is a 'tape-backed' (or 'd0t1') system for long-term archival storage, with 4 sizeable (300-700TB) disk caches to allow data to be conveniently staged for bulk recall. This is used by both the WLCG and local users at RAL (such as the Diamond Light Source[7] and the Centre for Environmental Data Analysis[8]). This is expected to be retained until mid-2020 at the very earliest, and no decision has been taken on its replacement. RAL has 36PB of data on tape, fronted by 2.2PB of cache disk.

The second is a 13PB 'disk-only' (or 'd1t0') element, intended to be used for high-speed access by the Tier 1's Condor[9] batch farm and other low-latency use cases. This storage provides 'permanent' disk storage, in that data will be retained until the user deletes it. It does not, however, provide the same level of data security that is offered by a system with tape backup. This CASTOR disk-only element is the one that has been discontinued at CERN and is being replaced by Echo at RAL.

This paper will first give an overview of the current Tier 1 CASTOR architecture and then discuss the modifications being made to RAL's CASTOR service in order to support this transition.

1.1 Current CASTOR Architecture

CASTOR is a highly database-centric system. Oracle databases are used for (among other things) the namespace, transaction handling, scheduling, permissions, and tape management. For our purposes, CASTOR databases can be divided into two categories – those that are replicated for each user (or group thereof), and those that are provisioned centrally for all users. The first category is known broadly as the 'stager', the second as 'central services'. Interfaces to the databases are provided by Unix daemons that run on CASTOR management nodes.

The RAL Tier 1 has four instances of the stager databases. One each is provided for the largest three LHC[10] experiments – ATLAS[11], CMS[12], and LHCb[13]. Another is shared between ALICE[14] and all other WLCG experiments – this is known as 'Gen'.

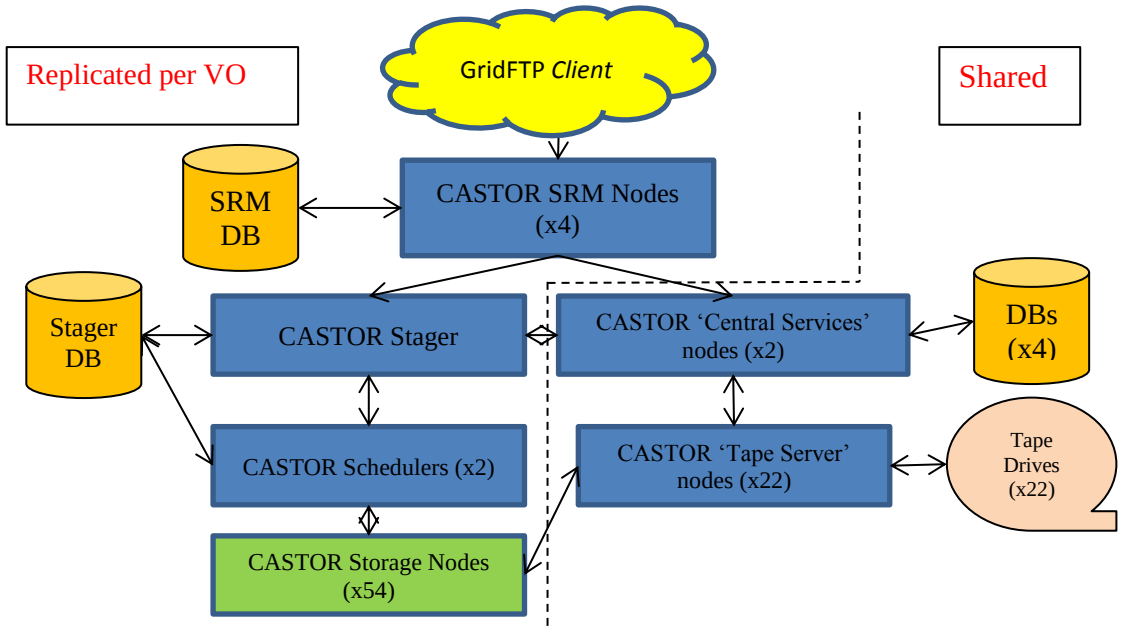
The role of the stager is to manage the state of files on disk – conceptually, they are being 'staged' in transit from disk to tape, even if in practice the use case is disk-only. Several processes, spread across three nodes, are currently required to do this. The Unix services that interact with the stager database include the 'stagerd' process that manages the storage nodes (also known as 'disk servers') and the files residing on them and a scheduling system that manages incoming and outgoing transfers between the storage nodes and the outside world. Storage nodes are owned by a given stager and cannot be shared. Also associated with the stager services are the processes that provide SRM[15] and xrootd[16] interfaces for external users. The CASTOR SRM also has an associated database – this is co-located with the stager.

Each stager instance has three management nodes for the stager and directly associated processes, and 2-4 more providing the SRM interfaces (depending on the amount of level of load seen on the instance). This adds up to 23 management nodes that used to run the stagers and SRMs.

The central services provide a variety of services to the system as a whole. The name server daemon ('nsd') is responsible for managing CASTOR's internal namespace, other processes manage authorization of management processes and oversee the tape drives. Two more management nodes are required to run these processes for the whole system.

The tape drives are shared – there are 22 in total, and any user is able to use any drive for reading or writing (although there are restrictions placed on the number each user can use concurrently).

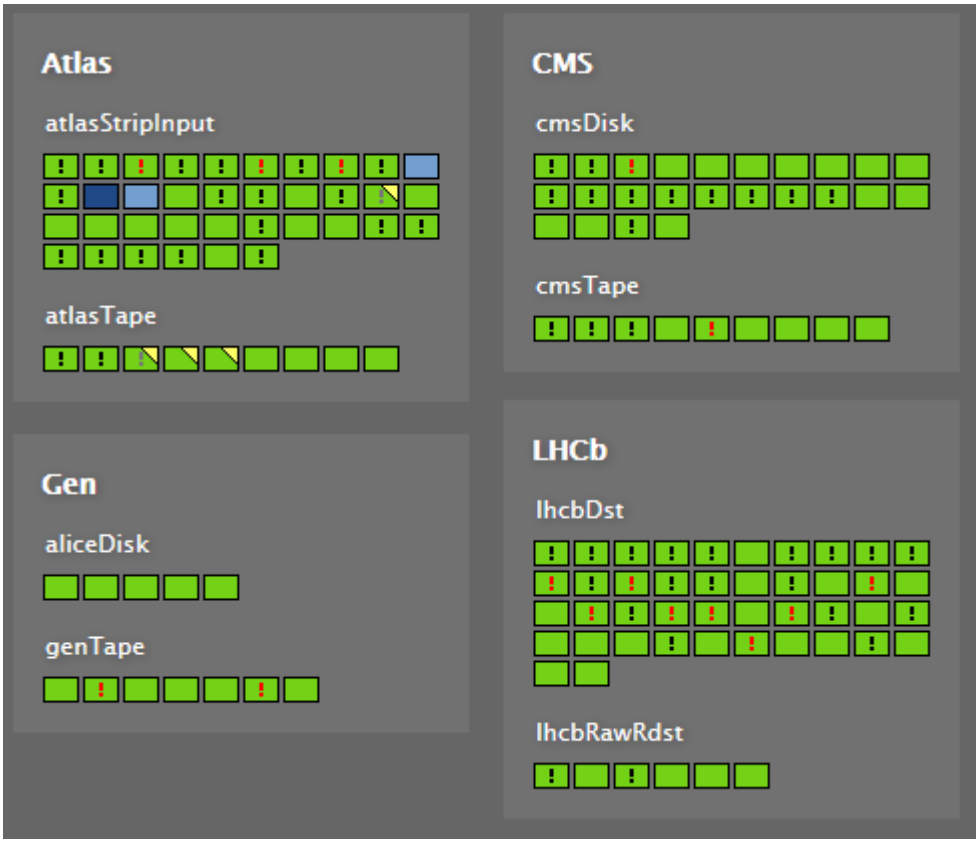
Figure 1: Current CASTOR Architecture



The databases on the diagrams above are provided by two Oracle RACs[17]. One is used to host the stagers for ATLAS and 'Gen' (ALICE and non-LHC WLCG). The other hosts CMS, LHCb and the central services. The transaction rate seen for each RAC in the order of 390Hz. This load is dominated (>95%) by disk-only operations – this is partly because files ingested to tape are larger (RAL's average file size on disk is 578MiB, average file size on tape is 1024MiB), but mostly because of the higher request rate and lower latency requirements imposed by the use case of the Tier 1 batch farm.

CASTOR disk storage is provided by storage nodes with capacities of 60-120TB, provisioned using RAID 6. The same kind of hardware is used for both 'disk-only' provision and the 'tape-backed' cache. Servers can be moved from one service type to the other without any hardware changes. These nodes use 10Gb network cards, but typically sustain a peak I/O performance of around 3Gb/s/node. After some investigation, we believe that the bottleneck is imposed by disk I/O issues stemming from the particulars of the RAID configuration. CASTOR has 137 of these storage nodes in total, 28 of which are used for d0t1 (caching in front of tape), while the remaining 106 are used for d1t0 (fast storage for batch).

Figure 2: Current CASTOR storage nodes



2. The planned approach

As noted above, the disk-only elements of the CASTOR service are due to be replaced by the new Echo system. This implies a drastic down-scoping of the service – as noted above, disk-only accounts for 77% of the storage nodes and 95% of the transactions. A policy of making no architectural changes and running the service exactly as before, except with no d1t0, would result in 25 management nodes (not counting tape servers) looking after 31 storage nodes and two RACs running at a fraction of the capacity of one. The overhead incurred by managing all of these infrastructure elements to support a much-reduced number of storage nodes would be unacceptably large. We therefore plan to re-implement the CASTOR service.

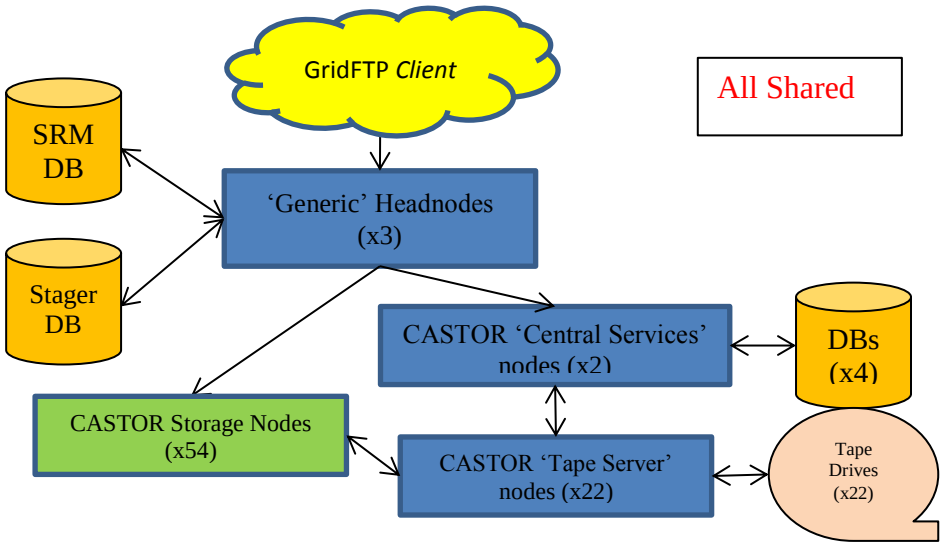
2.1 Project Objectives

The objectives of the CASTOR re-implementation project are as follows:

1. Avoid the loss of any data
2. Reduce the quantity of management nodes to a level more proportionate to the expected requirements
3. Reduce the amount of staff effort required to manage the system
4. Improve the reliability of the system

2.2 Target architecture

Figure 3: Planned future CASTOR architecture



In order to achieve these objectives, we intend to redeploy CASTOR using a new architecture, as shown in figure 3.

There are multiple changes contained within this modification:

2.2.1 Merging of CASTOR instances

The reader will note that the dashed line from figure 1 has been removed in figure 3. This signifies a fundamental change in how the service will be provisioned. Under the current configuration, 4 separate instances of the stager database exist. The new architecture proposes to use only 1. This will be a single instance of the stager shared between all users. This is possible because of the large portion (~95%, as noted above) of the CASTOR request rate that is accounted for by disk-only operations. This will also allow (indeed, require) us to retire one of the Oracle RACs used to support CASTOR's databases. We will also reduce the number of tape-backed disk caches that are used. Currently, we have 4, one for each stager instance. One stager instance can support more than one cache, but a single one is preferred, at least initially, in order to avoid the efficiency problems inherent in statically partitioning a resource that the demand for varies dynamically.

The process by which this will be done is simple. We will declare the instance to be down, and then the aliases used by the experiment will be re-pointed from the old management nodes to the new ones. Experiment-side functional tests will be performed, and then the new instance will be opened to the user. This migration from a separate stager instance to a single one will require an interruption to service, but we expect it to be a short one (~1 hour). Each user can be migrated separately.

We do not propose to migrate any data from the old user-specific stagers to the new shared one. This does not imply data loss. The migration will occur once the user has removed all of their 'd1t0' data from CASTOR, leaving (in theory) only the contents of the tape-backed disk cache in the stager (in practice, some data will inevitably remain due to inconsistencies between the storage service and the experiment's data catalogue, but large WLCG experiments view their own catalogue as authoritative – any data remaining will be 'dark' and not of use to the experiment).

2.2.2 Shift to virtualisation and 'cattle headnodes'

A new, 'blank slate', CASTOR instance offers a chance for other architectural improvements to be made. Currently, CASTOR stager management nodes (also known as 'headnodes') at RAL are provisioned on physical hardware and are non-redundant – various critical services run in one place and one place only, meaning that a single hardware failure can render the entire service inoperable. This is a flaw that we intend to address.

The new management nodes will be generic and failure-tolerant ('cattle' as opposed to 'pets'). This will be achieved by provisioning new systems such that all services run on all management nodes, as opposed to the current configuration where each node runs a bespoke subset of possible processes. Load will be distributed amongst these new nodes using HAProxy[18]. HAProxy also solves another problem - it can be configured such that if a process stops, this is automatically detected and requests no longer sent to that process, instead being diverted to another instance of it. We expect this to be of great use on occasions where it is necessary to restart hardware (such as for kernel patches).

Virtualisation technology will be used to provision the new management nodes. This is not a pure improvement on physical hardware in terms of reliability (it shifts the failure domain from the node level to the level of the hypervisor), but is expected to offer more flexible scaling of the service should it become necessary to add or remove management nodes. It is anticipated that 2 or 3 virtualised nodes should be sufficient to serve the new, merged stager instances, plus another 2 for the central services.

2.3 Possible problems

Two possible problems with the approach above should be noted.

The first is the possibility of contention between users inherent in offering a shared resource. RAL already sees some element of this – the drives used to write and read data to/from tape are a shared resource, and on occasion users will find themselves unable to interact with tape due to the actions of another user. The most worrisome possibility in the new configuration is for one user to see data that they recalled to the disk cache for later use flushed from the cache before they are ready due to another user attempting the same thing. In order to mitigate this the size of the cache will be relatively large (>1PB). Experience will tell us whether this becomes a significant issue in practice, and it will be possible for us to repartition the cache into static pools once more should it be required.

The second is potential difficulties in scheduling interventions that require the service to be taken down. Because the service would be used by many different communities, each with their own requirements and timetables, it may become very difficult to find a time that is convenient for all. Given the improved setup (redundant management nodes fronted by HAProxy that allow a single node to be taken down without an interruption to service), we do not expect this to be as frequent an occurrence as it has been in the past, but there will still be times when it happens. The factor that makes this tolerable is the fact that only tape access is being provided using CASTOR – tape access tends to be relatively orderly and non-time-critical, as opposed to the relatively 'chaotic' nature of WLCG disk-only usage.

3. The future of CASTOR at RAL

The authors note recent changes to planned tape provision at CERN, in particular the planned retirement of CASTOR in mid-2019 in favour of the Cern Tape Archive (CTA)[19]. At this point, no further development effort will be expended on CASTOR development from the CERN side.

At the time of writing, no decision has been taken on how to respond to this. The loss of support for CASTOR would certainly suggest the need to migrate away, but migrating to a whole new tape store will (almost inevitably) be a time-consuming process, in the order of years (this is unless the Tier-1 chooses to use CTA, which uses the same tape format as CASTOR and so requires no tape migration). We therefore are proceeding with this project on the basis that it is likely to have plenty of time to bear fruit in the form of improved service reliability and availability, and reduced staff effort in maintenance.

4. Conclusion

We have described how we intend to use the opportunity provided by the partial replacement of RAL's CASTOR system to reduce the service's operational overhead, particularly by desegregating the disk cache layer and replacing a large number of non-redundant management nodes with a smaller, redundantly configured set. We have also described the method by which we intend to reduce the scale of the CASTOR service without long downtimes for our users.

References

[1] Worldwide LHC Computing Grid – http://wlcg.web.cern.ch

[2] D. Britton, et al. *GridPP: the UK grid for particle physics*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **367.1897** (2009) 2447-2457B.

[3] G.L. Presti, et al. *CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN*. MSST. Vol 07 (2007), pp.285-80

[4] *CASTOR - CERN Advanced STORage manager* http://castor.web.cern.ch/castor/

[5] A. Dewhurst, et al. *The deployment of a large scale object store at the RAL Tier-1*, Journal of Physics *Conference Series **898*** **6** (2017) IOP Publishing

[6] S. Weil, et al. *Ceph: A scalable, high-performance distributed file system*, Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, 2006.

[7] The Diamond Light Source – https://www.diamond.ac.uk/Home/About.html

[8] The Centre for Environmental Data Analysis – http://www.ceda.ac.uk/about/

[9] D. Thain, T. Tannenbaum, M. Livny, *Distributed computing in practice: the Condor experience*. Concurrency and computation: practice and experience. **17.2-4** (2005): 323-356

[10] L. Evans and P. Bryant, *LHC machine, Journal of instrumentation* **3.08** (2008): S08001

[11] ATLAS Collaboration, *ATLAS: technical proposal for a general-purpose pp experiment at the Large Hadron Collider at CERN*, **CERN-LHCC-94-43**, http://cdsweb.cern.ch/record/290968

[12] CMS Collaboration, *CMS technical proposal*, **CERN-LHCC-94-38**, http://cdsweb.cern.ch/record/290969

[13] LHCb Collaboration, *LHCb technical proposal*, **CERN-LHCC-98-004**, http://cdsweb.cern.ch/record/622031

[14] ALICE collaboration, *ALICE: Technical proposal for a Large Ion collider Experiment at the CERN LHC*, **CERN-LHCC-95-71**, http://cdsweb.cern.ch/record/293391

[15] F. Donno, et al. *Storage Resource Manager Version 2.2: design, implementation, and testing experience*, *Journal of Physics: Conference Series **119*** **6** (2008) IOP Publishing

[16] A. Dorigo, et al. *XROOTD-A Highly scalable architecture for data access, WSEAS Transactions on Computers* **1.4.3** (2005)

[17] Oracle Real Application Clusters – http://www.oracle.com/technetwork/database/options/clustering/overview/index-086583.html

[18] D. Patterson, G. Gibson, and R. H. Katz. *A case for redundant arrays of inexpensive disks (RAID).* ACM Vol. 17. (1988) No. 3.

[19] HAProxy - The Reliable, High Performance TCP/HTTP Load Balancer http://www.haproxy.org

[20] S. Murray, et al. *An efficient, modular and simple tape archiving solution for LHC Run-3, Journal of Physics Conference Series **898*** **6** (2017) IOP Publishing

PoS(ISGC 2018 & FCDD)026