2nd ESCAPE WP2/DIOS workshop

9-10 December 2020

Indico Event

Participants (! = not registered)

Agustin Bruzzese !
Alba Vendrell
Aleem Sarwar
Andrea Ceccanti
Alessandra Doria
Bastien Gounon
Barkey Turk
Bernardino Spisso
Diego Ciangotinni
Daniele Spiga
Elena (?) ! Not registered
Fabio Hernandez
Federica Agostini !
Frederic Gillardo
Gareth Hughes
Ghita Rahal
Gino Marchetti
Lucia Morganti
Marek Szuba
Marcelo Vilac,a !
Mario Lassnig
Markus Elsing !
Martin Barisits
Nadine Neyroud
Patrick Fuhrmann
Paul Kramp
Paul Musset
Paul Millar
Pierre Chanial !
Raymond Oonk
Riccardo Di Maria
Rizart Dona
Rosie Bolton
Rohini Joshi
Simone Campana !
Stephane Jezequel
Tim Wetzel !
Tomasco Boccali
Trude von Richthofen
Vishambhar Nath Pandey
Xavier Espinal
Yan Grange

Dress rehearsal review

Pilot assessment

Experiment View (Wednesday 9, December 2020)

Rosie Bolton (Square Kilometre Array Organisation)

Workshop opening
Presentation Link

Q&A:
VnPandey: How long should the logbook be ?
Rosie Bolton: The most information as possible. Idiot proof. What went wrong what went well - but with a summary at the end that can be included in the deliverable.

LOFAR
Speaker: Yan Grange (ASTRON, the Netherlands Institute for Radio Astronomy)
Presentation Link
Vishambhar Nath Pandey (ASTRON) , **Yan Grange** (ASTRON, the Netherlands Institute for Radio Astronomy)

Q&A:

Raymond Oonk : Single file transference or Multi-file transference?
Yan : Multiple in Parallel.

Agustin: Further intentions on using a non deterministic RSE? We ll be happy providing you some help on that.
Yan : We want to try a non-deterministic RSE. We can talk afterwards.

Simone : Use case from LOFAR is to be equivalent of Tier0 at LHC so people like Mario and Martin (RUCIO experts) have experience
Do you publish and use downtime? For example Rucio could use that to stop sending data there.
Xavi : Not yet but this is the plan. Use GOCDB+ CRIC as a main feeder for RUCIO to know about unavailable RSEs. Also in terms of ticket and site interaction the idea is to have a look at GGUS (KIT). Both GoCDB and GGUS are part of EGI and can be part of EOSC.

FAIR

Speaker: Marek Szuba (GSI)
Presentation Link

Q&A:
Martin Barisits: The rucio upload client is not optimized to do parallel upload.
Something to run in the background for token issuing - doing this on the fly causes overhead.

LSST

Speaker: Bastien Gounon (CC-IN2P3)
Presentation Link
Bastien Gounon (CC-IN2P3) , Fabio Hernandez (CC-IN2P3) , Ghita Rahal (IN2P3/CNRS)

Q&A:
VnPandey: What does the image represent ? (slide 11)
Bastien: It's actually a view from the sky.
VnPandey: Some error messages are not completely clear, and need a closer look (slide 7)
Bastien: I use the debug option in the Rucio client (slide 7)
Riccardo: Pandey you couldn't see an error because the Rucio server was manually down during this 5 minutes time window (slide 6-7)

Rosie: You said you had to relaunch the failed upload, did you have to do that manually or was it done automatically?
Bastien: We had to do it manually. In order to force the upload. Would be great to be done automatically.
Martin: [Rucio] does a few tries automatically and then stops.
Paul Millar (in the chat) : Were you able to understand which RSE was failing when an upload failed?
Bastien : the only RSE we were using was dCache at CC-IN2P3

EGO/VIRGO

Speaker: Dr Pierre Chanial (is going to do the presentation instead of Elena Cuoco)

presentation link

Elena Cuoco, Dr Pierre Chanial (European Gravitational Observatory)

Q&A

Simone : About proprietary data: Everyone is interested in this used case. Ideally it should be done with the click of a button to pass from proprietary to opendata. ESCAPE seems to me the perfect scenario to start investigating about that.
Andrea : This was not in the scope for the current DR, but definitely, it was planned for next year. Protection and data separation is not trivial. Rucio needs to understand this. We already started a proposal on how this could be achieved.
Xavi : We did some work around already for CMS some months ago. Regarding proprietary data.

Martin: about events. You should really used events for this use case because if not you are hammering the Rucio instance
Pierre: How do I do that ?
Martin: I will send you a code snippet
Riccardo: Me and Rizart are following up on that. Need access to CERN's ActiveMQ broker
Pierre/Martin: Will I get all the events for all scopes? Yes, everything then you need to filter for what you are interested for.

SKA

Speaker: Rohini Joshi (SKA)

Presentation Link

Rohini Joshi, Rosie Bolton (Square Kilometre Array Organisation)

Q&A

Martin: Really interesting test. First time we had a data from that side. Changed rucio algorithm to not require so much memory
Rosie : Is a memory or an algorithm problem ?
Martin: This daemon tries to put every replica in memory. The new algorithm does it by groups of replica so taking constant memory, which means better scalability to large datasets.

Stephane: Do you plan to include RSEs from outside Europe - e.g South Africa, Australia?
Rohini: That's something it's on the list. Depend on the people's time.

Xavi: The relation with Australia slowed down during the pandemic. Hoping to relaunch it soon.

ATLAS

Speaker: Stephane JEZEQUEL (LAPP)
Presentation Link

Q&A

Xavi: I have some questions regarding the conversion from the Atlas instance into the ESCAPE project.

CMS

Speaker: Diego Ciangottini (INFN)
Presentation Link
Daniele Spiga (INFN) **, Diego Ciangottini** (INFN, Perugia)

Q&A

Xavi : Thanks Diego. Basically, you've touched a very interesting topic [token use]. So, I'll probably give the word to Rizart. How has more knowledge into the X509-free services.
Rizart: This is something I've recently started to looking at. The X509-free Rucio is just a matter of looking at the configuration. For ESCAPE It's just a matter of configuring the server to accept tokens. This will definitely start next year as soon as possible.

Andrea: Regarding the WP5 workshop: Has been postponed to January; Encourage to join - this will be discussed there.

MAGIC

Speaker: Agustin Bruzzese (PIC)
Presentation Link

Q&A

Xavi: Could be a good test for similar experiments.
Stephane : How easy is to get ..
As far as I know, it is a collaborative work, you need to configure Rucio and RSE's to deal with embargoed data. For now we have a workaround to do it .

Stephane: Is it documented somewhere?
Andrea: We wrote a proposal on how to write embargoed data on the datalake

Paul Millar (by chat): Please note that the "Xavi work-around" (where there is a dedicated RSE for embargoed data) causes problems, as Rucio does not know of this restriction, and tries to replica other people's data there.
I would recommend against using this work-around as it causes other issues.
Xavi (on chat): I do not see the problem to use this approach as a test before we have an harmonised implementation for embargo data. Such an RSE can work "hidden" .

CTA

Speaker: Frederic Gillardo (LAPP)
Presentation Link
Frederic Gillardo, Berkey Turk (LAPP/CNRS)

Q&A

Marek : We need a functioning mechanism in our data lake.
Andrea: I ‘ve a suggestion on the apache server you are currently using(side 2). I suggest you take a look at stand alone web application (non-apache) as we have in the STORM-WEB-DAV
Frederic: Is already package as a docker image with OIDC support and checksum problem is fixed by Barkey.
Andrea ; Yes indeed. We can provide it and let us know . We can discuss this offline.

Paul : The out of memory error comes from the amount of connections that were happening at the same time. It’s quite difficult to debug it. However we can work on it.
There are sites that don't want to install a big storage system and we’re planning on addressing that with German Helmholtz Association. We’re looking at Apache as an easy way of allowing access. Hepix talk where they talk about that here : https://indico.cern.ch/event/898285/contributions/4041954/

Experiment's wrap up: identified challenges, prototype plans focus

Embargo data
Tokens
FTS/network throttling multi-VO
RUCIO outside WLCG
RUCIO event notification mechanism - in place.
Downtime/ ticketing GOCDB/GGUS

Paul: FTS Multi VO part we have seen with GSI use case, They have a rather low power disk server. Is it realistic to have a low storage endpoint.

Xavi: [...] From my understanding: Yes. [...]

Rosie: Will be good idea in the ESCAPE project to bring new sites and projects ___
Martin: We should maybe try the other injection workflow to put the data on RSE using something else than the Rucio Client that was not done for big data injection. But we can also improve the data client to inject more data
Rizart: Event rucio notification mechanism: This is already in place, the events are being produced by the rucio hermes daemon. The mechanism is in place, if anybody wants to consume these events, mail me.

Paul Millar: in Dcache we have an event system also, we have to systems one for the system/site level, another one more web oriented addressed to user. Would it be interesting for Rucio to (didnt follow after that)
Martin: Yes we could be able to do something like on demand event in Rucio but its not priority now.
.Pandey: What are the minimum HW and SW resources are needed; How to handle the becoming RSE.
Raymond from chat: AT SURFsara we implemented a wrapper around the dCache API that can also be used to listen to events https://github.com/sara-nl/SpiderScripts Its called ‘ADA’

Services and continuous testing

Stephane JEZEQUEL (LAPP)

RUCIO

Speaker: Dr Riccardo Di Maria (Cern)
Presentation link

Q&A

Stephane: After the issues you saw during the FDR, do you think it will be ok for the coming challenges?
Riccardo (answered, but I didn't get everything). I'm almost certain that the current set up will be enough for the upcoming tests.

Rohini: The helm chart that has been deployed can they be shared ?
Riccardo: They are the one from Rucio (slide-12)

AAI

Speaker: Andrea Ceccanti (INFN)
Presentation link

Q&A

Rosie: about high availability IAM
Andrea: I don't see how we can have any big incident by next week. What we need to have is another k8s cluster, holding the IAM backup instance.

Xavi: We discussed this last week in wp2 meeting; No official FDR but there will be a golden time slot for this activity next week.

Paul Millar: Will the redirector be also redondant?
Andrea: Yes

Monitoring

Speaker: Rizart Dona (CERN)
Presentation link

Q&A

Riccardo: About the last sentence from Rizart; Until we recreate the all cluster there is no way to monitorize.
Rohini: the Rucio stat dashboard about DIDs per scope,
Rizart: This run every 10 minutes .. As more data you upload, and more DIDs, this is going to be slower to get the data back. Depends on how rucio query the database, still working on that.

Ghita: I missed the point about what you said about the Rucio dashboard
Rizart: The basic monitoring comes by the rucio events, this do not reflect upload operations done to the datalake.

Discussion: Continuous testing frameworks

Rizart Dona (CERN)**, Rohini Joshi** (SKA)
Presentation link

Q&A
Andrea: On the slack webhook: Probably you could also have a rocket chat (*).
It looks Rucio focussed as for IAM it would be more about … So a good thing would be to have a status page with either up and down for all services beginning with the core services

(*)Yan from chat:I have a script that posts a DM to rocket chat. Can share the code with you (may be on some git actually :D)

Rohini: That would be great, thanks Yan.
Rosie: Suggests; Public so you don't need a grafana account.
Rizart: If you use some external tool like Andrea was talking about it it would be good. I don't know if we can have persistently open without accounts.

Site perspective (Thursday 10, December 2020)

CC-IN2P3

Speaker: Ghita Rahal (IN2P3/CNRS)
Presentation link
Ghita Rahal (IN2P3/CNRS), **Paul Musset** (CNRS)

Q&A
Ricaardo: Problems faced regarding tokens-based Authentication.
Ghita Rahal: We are are unable to install the service to be able to validate the token, simply because there's no client deployment (No root access).

Paul: You are using the root protocol thought xcache

Paul Musset: We haven't tried xrootd 5 yet so we didnt try token with the xrootd protocol. Token with webdav works even in xrootd4.
Ghita: Token on XCache was working only with webdav. We tried only webdav both through the XCache and directly to dCache
Riccardo from chat: Indeed and in R5 it is embedded within xrootd

GSI

Speaker: Paul Kramp (GSI)
Presentation link

Q&A
Xavi comment: Great work by the GSI team, good to see that this site worked so well with a single disk running on a VM, but good also to see it evolving into the prototype phase

Paul Millar: Did you open a ticket to the slack expert developers about this undefined state
Paul Kramp: We didn't,

Xavi: These might be caused by write/read locks on a thrashed disk…
Marek to Xavi on chat: That's my suspicion too but the (somewhat funny) thing is, they start during a load spike but persist even after the server has calmed down. Which is why we like to use the term "entering undefined state" here.
Xavi to Marek on chat: mmm… once you have thread mixing rwlocks, they can enter into a EBUSY state so lock no cannot be released… Certainly Andy will give you the right answer on that!

Paul Millar: When creating a new RSE we could use here to do the replication from the old one to the new one rather than doing it manually and injecting it.

LAPP

Speaker: Frederic Gillardo (LAPP)
Presentation link

Q&A

Paul Millar: Answer to the last part, Desy have a DMZ and services in the DMZ are considered kind of… so its easier to have
At dasy we're running dCache and we have the same issue you had. It's a well understood problem by the security team. In Wlcg, there is a list of subnet and not a list of ip. To not have a continuously changing list of IP.

Frederic: This is what we did at the beginning. We provided a list of IPs and ? And I remember someone from SKA tried to connect and I realized it was not open for (all the sites?)

Paul: Your dCache is 10TB(slide-2). Do you plan to increase that?
Frederic: For now the server is 40TB but some disks are busy. I can clean them
Paul: You can expend by adding a new server

Marek: For a system like the data lake having a white list of IP address….. Instant change recently because of the migration… Another thing: Most of the clients will mostly be talking from home but it's not guaranteed that someone wi ?
GSI side, we are somewhat lucky because even is security people are quite strict, they are also quite comprehensive.

Discussion

Simone: about the security Thing, there are is opportunity in escape. For LAPP as a WCG tier 2, it should be already a problem.

Frederic: LAPP administrator can not be sure that what I do is really up to standard with security

Simone: In our community here (high energy and nuclear physics and Astroparticle Physics) there is an opportunity to create this network of trust. What we realized in WCG a long time ago, is that a lot of those security aspects are a bit technical, but mostly they have to do with policy and trust. If we can create a larger trust in a larger community then we can achieve a security infrastructure, even for future production needs. Romain Wartel, responsible of WLCG security try to push that a lot in other communities. But there is some push back. But if you do that early it will help your life in the future.

Ghita: This is great comment and will lead us in the direction to set up EOSC.
Xavier: my comment was about Simone's point. Last week I met with Romain and we basically started to discuss about that, and the plan was that Romain will join one of our meetings and will explain a little bit.

Riccardo : Cern security team is scanning our VM (said during the pause after a question from Ghita)

FDR wrap-up

Speaker: **Rosie Bolton** (SKAO)
Presentation link

2021 Program of work

Speaker: Xavier Espinal (CERN)
Presentation link

Task Lead Round Table/Discussion: main items collected from the workshop, some identified goals for 2021

Task 2.1 Data Lake Infrastructure and Federation Services (Riccardo)

- Quota of Data lake (Exabyte scale as proposed in Proposal)
- Real Use Cases

Task 2.2 Data Lake orchestration services (Paul)

- Continue work with experiments/facilities to build QoS work-flows and description
  - Review existing contributions: ATLAS, CTA, SKA, GSI …
  - Add new contributions: CMS, Virgo, …
- Following and support Rucio developments:
  - Supporting new "VO policies": deployment, testing, suport VO usage.
  - Exercise QoS transitioning: Rucio + FTS + CDMI as place-holder for future development.

- Help restructuring physical storage:
    - Reorganising data location on the storage service.
    - Not (only) a QoS issue, but QoS has an impact.
    - Include novel storage types: tape storage, EC, possibly volatile/opportunistic storage.
- Exercise more work-flows: Improve our practical understanding of QoS and (possibly) catch problems we discover by "actually doing it".

Task 2.3 Integration with compute service (Yan)

- Project based data access (If I understood correctly)

Task 2.4 Networking and monitoring (Rosie)

- Test running (Networking)
- Cloud Storage (SKA is using commercial Cloud Storage)
- ESCAPE external partner involve

Task 2.5 AAI: Authentication and Authorization (Andrea)
Presentation link

Q&A
Rosie: End day of the project?
Xavi: The official end date is still 06/08 but we asked for a +6 months extension

Marek: What do you mean by external compute resource ? Comercial cloud ? External to the data lake?
Xavi: It would be great to have some HPC resources, some comercial cloud resources, but it would be good that some sites provide external compute resources. Example: They have some cluster that runs kubernetes and they expose that so this could be also seen as an external thing.
Marek; in that case there is INFRAEOSC07 where there will be fedecloud (?), that we can use as one this external ressource. Where DESY is involved.

Task 2.1

Riccardo: with the fdr, we saw some issues. With pilot phase, we are good. But we can improve a bit, a lot for prototype phase.
We could improve the whole quota for all RSEs.
We Could also identify new workflows that some experiments could include in his experiments.

Xavi: We cannot of course do Exabyte-scale. The goal is to see if we can scale it enough. LHC experiments are using ESCAPE at a test. For new sciences/experiments, they are using ESCAPE as a way to look for a useful computing model.

Rosie: Write down a pathway to Exascale - maybe as a public memo. Experiments have different needs/requirements.

Ghita: 1. We tried to do our best as a site to increase our data storage volume but it's out of possibilities: We'll not be able to increase. 2. Thinking about a service like xCache that we tried to test; How can be tested really functionally for experiment use cases? How we can test some of the services that are not RUCIO, FTS, etc. in the context of new experiments that have not experienced the normal data lake.
Riccardo:

Xavi: We should really work with WP5 to have more workflow exemple of how a cache can be used.

Ghita: We have to find use cases to test cache

Riccardo: For CMS, we have a lot of tier2 where basically using tier2 as cache. Maybe do temporal caches.

Task 2.2 QoS

Paul Millar: Rucio as evolution with VO/QOS. QOS-transition where you allow a storage system to facilitate transition from one medium to another. Reorganise the RSE for experiments/user to have separate spaces. Have storage with different QOS type(tape, volatile opportunistic storage)

Xavi: QOS-transition : is it possible having a test-bed using that ?

Paul: For EOS and.. Storm? this is not a problem, theoretically it should work. If it doesn't work we may need to go to plan B.

Rosie: Do we have access to tape storage?
Paul: We have a couple of site (WLCG tier1 site) have some tape access like PIC
Rosie: Is it clear who leads ..

Ghita: CC-IN2P3 could give access to tape. QOS-rule are not defined by medium but is a common rule that the site will choose how to set at their level ?
Xavi: yes

Yan: Different ways to access the datalake ? For example, use case from someone that doesn't want to know all the inside of the datalake.
Xavi: yes collaboration with wp5

Rosie, Xavi and Andrea: Discuss about the 21st of January authentication workshop. No problem to be fairly less technical and more focused on use cases (Our WP2 WP5 meeting). Also could be a good idea to combine workshops. This is going to be discussed offline.

Task 2.4 Networking

Rosie: Networking : test running. Thanks Rizart and monitoring team.
Could be nice to do long haul data transfer, for example with AARNET., extend the ruio instance outside Europe

Can Rucio/FTS be configured to fill a dedicated link without scheduling or is more control required?

Martin: About metadata: Nice to test next year. About large scale transfer tests: We probably have to check with the FTS people.

Xavi: About the cloud storage: Strange a bit the collaboration with the cloud storage __

Raymond: Testing plans for next year - high throughput links (400 gb/s) involving SurfSARA, overlap with CS3MESH4EOSC project too.
Rosie - will co-opt Raymond into SKA Regional Centre Data Logistics working group

Need to remember that not all experiments were yet able to participate in the FDR - we need to help these get online. (KM3NET at least…)

Also we have interest from additional European sites - e.g. Sweden

AAI

Andrea Ceccanti
Presentation link

Marek: Have you confirmed that native xrootd with token authentication is still not possible with xrootd5?
Andrea: There are still some implementation aspects that need to be sorted out on token support.

Ghita: Question about namespaces where you have the ESCAPE (d. 4). Acces under ESCAPE = ESCAPE view?
Andrea: In any case this is ESCAPE focus.

Martin: Wondering about the __ data.
Andrea: The name space structure was easily achievable but the embargoed data is a bit more difficult. The proposal have to be discussed if it is too difficult. ACTION - bring a discussion on embargoed data this into the next WP2 meeting

List of action items/ wish list

1. Communicate downtime
2. Timeouts on storage - upload fail (Yan)
3. Network monitoring plots (Yan)
4. Non-deterministic RSEs - have more available to help upload (Yan, Agustin, Rohini)
5. Off-line rucio token renewal to avoid this as a bottleneck (Marek)
6. Rucio events - event driven notifications (Pierre)
7. Have rucio rules with a start time too - future rules
8. Dashboards - have "No DATA" not red - neutral
9. Token based tests
10. embargoed data (Diego (CMS), Pierre (VIRGO), SKA, MAGIC)
11. Improvements to upload functionality
12. FTS throttling / configuration to support RSEs of different scales
13. IAM monitoring results pushed into CERN DB onto live dashboard
14. High Availability IAM deployment - proof of concept - for next year
15. Rocket chat web hook? / test results into rocket chat
16. High-level status page
17. Security / open access / do we need DMZ for public services? Do we need common method? (future work)
18. Write down a pathway to Exascale (Rosie)

Workshop Tidy-up

1. get all presentations on indico DONE!
2. make videos for each presentation
3. Put presentations and video links on the WP2 wiki (?) and on indico - notify participants
4. Tell WP6 (comms) - write up "blog post" entry