
Einstein Telescope E-Infrastructure Board

Div1 : Software, Frameworks and Data Challenge Support

Andres Tanasijczuk

EIB Workshop, Aachen, Germany

9 - 10 March 2023

Division general info

Chair: Andres Tanasijczuk (andres.tanasijczuk@uclouvain.be)

Liaison with OSB Div10: John Veight

Mailing list: et-eib-div1@ego-gw.it

Subscribe at <https://mail.ego-gw.it/mailman/listinfo/et-eib-div1>

Wiki page: <https://wiki.et-gw.eu/EIB/SoftwareFrameworks/WebHome>

GitLab group: <https://gitlab.et-gw.eu/eib/div1>

Division mandate

Division 1: Software, frameworks, and data challenge support

Define the software frameworks for ET computing workflows, the middleware for infrastructure, workload and data management. Develop software quality best practices and support their adoption with training and enforcement policies. Support code development in all computing domains. Provide computing support for mock-data challenges.

- Collaborate with OSB and ISB to define the data formats (both internal and for public release) and organized data processing workflows
- Support the development of the tools for the operation of the telescope,
- Coordinate the development of common infrastructural tools and frameworks for the data-analysis
- Support the operation of large-scale computing campaigns
- Develop policies and best practices to ensure software quality, and encourage/enforce their adoption
- Organize a continuous training programme for both developers and users

Organization of this session

- Part 1: Recent activities to support the Mock-Data Challenge
(same presentation I gave on Tuesday in the ET Monthly meeting)
- Part 2: Proposed (short) list of tasks on which, I think, we should start working on next
Open for discussion and do not hesitate to contradict me

Part 1

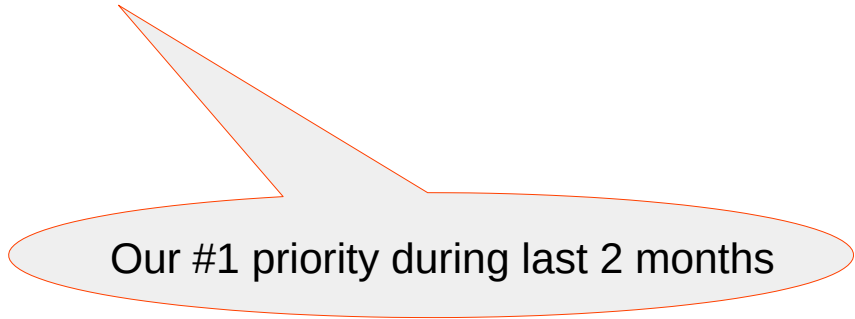
Recent activities to support the Mock-Data Challenge

Division mandate again

Define the software frameworks for ET computing workflows, the middleware for infrastructure, workload and data management.

Develop software quality best practices and support their adoption with training and enforcement policies. Support code development in all computing domains.

Provide computing support for mock-data challenges.



Our #1 priority during last 2 months

Mock-Data Challenge support

Provide access to MDC input data

Input data consists of 2 samples: noise-only and with CBC signal injection

1.24 TB : 620 GB per sample (155 GB x 4 channels)

1300 gwf files, per channel, per sample, of 2048 seconds each (total equivalent of 1 month)

We don't know on which resources will users run their analyses

Could be on their laptop, on a Virgo computing centre, on another cluster

We didn't receive any request to provide computing resources for the MDC, so far

In case we do, some resources have been secured at INFN-CNAF

Requirements:

Easy to distribute to any computing resources

Publicly accessible (at least for this first MDC)

Possibility to download files from a web browser

Make life less difficult for new users

MDC input data distribution

Decided to use same strategy and infrastructure as Virgo/LIGO, without authentication

Open Science Data Federation* (OSDF) infrastructure from Open Science Grid (OSG)

Service designed to support the sharing of files staged in autonomous “origins”, for efficient access to those files from anywhere in the world via a global namespace and network of caches.



Resources in EU provided by institutions affiliated to Virgo

- Virgo “origin” at UCLouvain
- Caches at PIC, IN2P3, CNAF, Amsterdam

Added a new “origin” for ET at UCLouvain

* Also known as StashCache

Virtual Organizations in OSG/OSDF

Each supported experiment/collaboration is called a “virtual organization” (VO) in OSG

Nothing to do with virtual organizations in grid computing

LIGO is one of these VOs, Virgo is not (Virgo is “hiding” under the LIGO umbrella)

OSG accepted to add ET as a supported VO in OSDF

“The entire global collaboration of GW observatories analyzes their data together”

Virtual organizations and OSDF resources are registered in [OSG topology](#) (yaml files in github)

Each virtual organization exports unique namespaces in the global OSDF namespace

Given the namespace, the VO can be unequivocally determined

Namespaces can be private or public

LIGO VO: `/user/ligo` (private), `/gwdata` (public) ET VO: `/et-gw/PUBLIC` (public)

Each VO defines a list of allowed “origins” and caches for each of its namespaces

ET.yaml file in OSG topology

<https://github.com/opensciencegrid/topology/blob/master/virtual-organizations/ET.yaml>

[... skipped lines ...]

```
39  OASIS:
40    UseOASIS: true
41    OASISRepoURLs:
42      - http://hcc-cvmfs-repo.unl.edu:8000/cvmfs/et-gw.osgstorage.org
43  DataFederations:
44    StashCache:
45      Namespaces:
46        - Path: /et-gw/PUBLIC
47        Authorizations:
48          - PUBLIC
49        AllowedOrigins:
50          - UCLouvain-ET-OSDF-Origin
51        AllowedCaches:
52          - ANY
53      DirList: http://et-origin.cism.ucl.ac.be:1094
```

OSDF “origins”

OSDF “origins” can be configured to make data private (authenticated “origins”) or public (unauthenticated “origins”), and can control the rules of sharing

Authenticated or unauthenticated: must be consistent with exported namespace being private or public

Authenticated (unauthenticated) “origins” give read access to VO allowed caches (everyone)

“Origin” admins can extend these default sharing rules, in particular for write access

OSDF “origins” are essentially [XRootD](#) servers with special configuration and additional supporting services

Authorized clients can interact with an “origin” server using XRootD client commands

List directory contents: `xrdfs xroot://<origin-fqdn>:<port>/ ls <namespace>/<subdir>`

Download file: `xrdcp xroot://<origin-fqdn>:<port>/<namespace>/<subdir>/<file> .`

Download directory: `xrdcp -r xroot://<origin-fqdn>:<port>/<namespace>/<subdir> .`

OSG provides instructions for [installing XRootD client](#) from their software repositories

Should already be installed in most computing centres (for sure in WLCG and IGWN)

Virgo & ET OSDF “origins”

Virgo “origin” was installed in a dedicated 100 TB storage server at UCLouvain

Private “origin” (https) listening on port 1095 (IPv4 and IPv6)

Uses Grid Security Infrastructure (GSI) protocol for authentication

Only Virgo/LIGO OSDF caches and Virgo data transfer operators are authorized to read the “origin”

Only Virgo data transfer operators are authorized to write into the “origin”

FQDN: `ingrid-se09.cism.ucl.ac.be`

ET “origin” service running on same Virgo “origin” storage server at UCLouvain

Runs in a docker container (provided by OSG)

Isolated from Virgo “origin” service - there is always the risk of a human mistake (from myself) though

Public “origin” (http) listening on port 1094 (IPv4 only)

No authentication: everyone has read access to the “origin”, from any host (except from writing host)

Only I am authorized to write into the “origin” (and from a certain host from where authentication is enforced)

FQDN: `et-origin.cism.ucl.ac.be` (DNS alias for `ingrid-se09.cism.ucl.ac.be`)

OSDF caches

OSDF caches provide a distributed data access layer to reduce wide-area network consumption, load on the data “origins”, and latency of data access

OSDF caches can also be configured as authenticated or unauthenticated

Each cache defines a list of allowed VOs for which they are willing to cache data

Virgo caches in EU allow all VOs (and ET VO allows all caches)

Authorization rules in authenticated caches are defined in VO namespaces

OSDF caches are also essentially [XRootD](#) servers with special configuration and additional supporting services

CVMFS

[CernVM File System](#) (CVMFS) is an HTTP-based file distribution service used to provide software and (small) data files in a fast, scalable, and reliable way

- Implemented as a POSIX read-only file system

- Files and directories are hosted on standard web servers (known as CVMFS Stratum-0) in so-called CVMFS repositories

- On client side, CVMFS repositories are mounted in the universal namespace `/cvmfs`

- Files are downloaded on demand and aggressively cached (squid proxies)

- Developed to deploy software for HEP collaborations on the [Worldwide LHC Computing Grid](#) (WLCG)

- Enhanced by OSG to distribute data files stored in OSDF

 - Supports authentication for private data (X509 proxies, tokens)

OSG provides instructions for [installing CVMFS client](#) from their software repositories

- Should already be installed in most computing centres (for sure in WLCG and IGWN)

CVMFS in GW community

CVMFS is used by LIGO, Virgo and Kagra (LVK) to distribute software and data

Managed by OSG, CVMFS server in the US

OSDF namespace `/user/ligo` is published in `/cvmfs/ligo.osgstorage.org` (private)

GW community is the main user of authenticated data in OSDF

OSDF namespace `/gwdata` is published in `/cvmfs/gwosc.osgstorage.org` (public)

Data from [Gravitational Wave Open Science Center](#) (GWOSC)

IGWN* software is published in `/cvmfs/oasis.opensciencegrid.org`

OSG created a (public) CVMFS repository for ET data

OSDF namespace `/et-gw/PUBLIC` is published in `/cvmfs/et-gw.osgstorage.org` (public)

MDC data available in `/cvmfs/et-gw.osgstorage.org/et-gw/PUBLIC/MDC1/`

May need to request the mounting of this CVMFS repository in computing centres that are not using auto mount

ET OSDF “origin” web server

Installed a web server on the ET OSDF “origin” for easier download of files

User doesn't need to install any software, as opposed to other access/download methods

<http://et-origin.cism.ucl.ac.be/>



Welcome to the ET OSDF Origin web server

This page contains instructions on how to download the input data files that were generated for the ET Mock Data Challenge 1, using this HTTP web server. For other ways of accessing or downloading the data, see the [EIB Div1 wiki](#) (ET credentials needed).

The input data files are grouped by channel (E0, E1, E2 and E3). There are 1300 .gwf files (155 GB) per channel.

Below are the links to the complete list of input data files:

With signal injection

- [E0 channel](#)
- [E1 channel](#)
- [E2 channel](#)
- [E3 channel](#)

ET OSDF “origin” web server

Installed a web server on the ET OSDF “origin” for easier download of files

User doesn't need to install any software, as opposed to other access/download methods

<http://et-origin.cism.ucl.ac.be/>

The screenshot shows the ET OSDF Origin web server interface. At the top, there is a header with the ET logo and the text "Welcome to the ET OSDF Origin web server". Below this, there is a section titled "Input data files for ET Mock Data Challenge 1 E0 channel". A callout bubble points to the first file in the list with the text "Click on a file to download". On the left side, there is a sidebar with a list of channels: E0 chan, E1 chan, E2 chan, and E3 chan. A red arrow points to the E0 chan link.

ET EINSTEIN
Welcome to the ET OSDF Origin web server

ET EINSTEIN TELESCOPE
Input data files for ET Mock Data Challenge 1
E0 channel

← Back to main page

Download list with all files

No.	File Name	Size
1	E-E0_STRAIN-1000000000-2048.gwf	122M
2	E-E0_STRAIN-1000002048-2048.gwf	122M
3	E-E0_STRAIN-1000004096-2048.gwf	122M
4	E-E0_STRAIN-1000006144-2048.gwf	122M
5	E-E0_STRAIN-1000008192-2048.gwf	122M
6	E-E0_STRAIN-1000010240-2048.gwf	122M
7	E-E0_STRAIN-1000012288-2048.gwf	122M
8	E-E0_STRAIN-1000014336-2048.gwf	122M
9	E-E0_STRAIN-1000016384-2048.gwf	122M
10	E-E0_STRAIN-1000018432-2048.gwf	122M

- E0 chan
- E1 chan
- E2 chan
- E3 chan

ET OSDF “origin” web server



Welcome to the ET OSDF Origin web server

This page contains instructions on how to download the input data files that were generated for the ET Mock Data Challenge 1, using this HTTP web server. For other ways of accessing or downloading the data, see the [EIB Div1 wiki](#) (ET credentials needed).

The input data files are grouped by channel (E0, E1, E2 and E3). There are 1300 .gwf files (155 GB) per channel.

Below are the links to the complete list of input data files:

With signal injection

- [E0 channel](#)
- [E1 channel](#)
- [E2 channel](#)
- [E3 channel](#)

Without signal injection

- [E0 channel \(noise-only\)](#)
- [E1 channel \(noise-only\)](#)
- [E2 channel \(noise-only\)](#)
- [E3 channel \(noise-only\)](#)

Parameters

- [Input parameters files](#)
- [Input lists files](#)

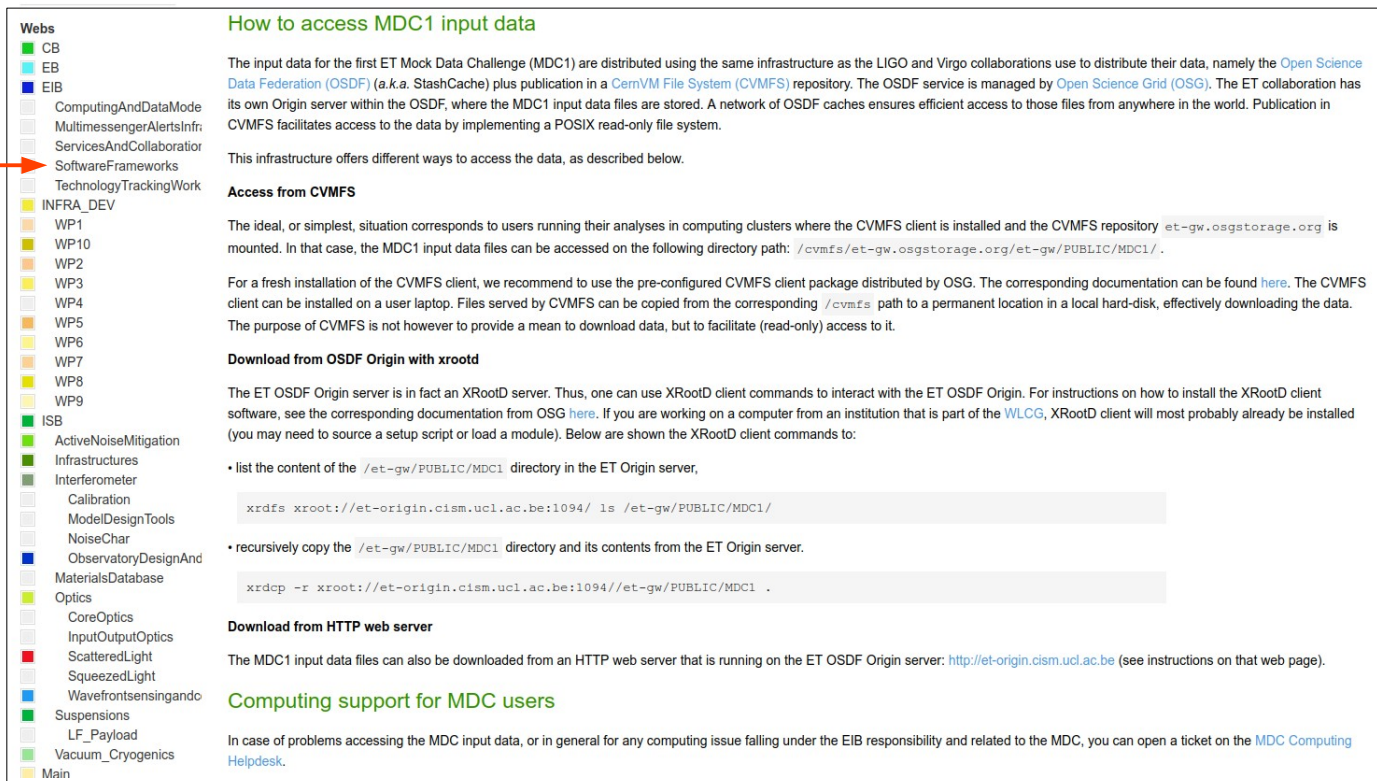
To download all the files with signal injection for a given channel, execute the following command in a terminal replacing `<channel>` by the corresponding channel (E0, E1, E2 or E3):

```
wget -e robots=off -N -r -np -A gwf http://et-origin.cism.ucl.ac.be/MDC1/<channel>
```

Instructions how to download all files
for a given channel with wget command

Instructions on how to access the MDC input data available in EIB Div1 Wiki

<https://wiki.et-gw.eu/EIB/SoftwareFrameworks/WebHome>



Webs

- CB
- EB
- EIB
- ComputingAndDataMode
- MultimesengerAlertsInfr
- ServicesAndCollaborator
- SoftwareFrameworks
- TechnologyTrackingWork
- INFRA_DEV
- WP1
- WP10
- WP2
- WP3
- WP4
- WP5
- WP6
- WP7
- WP8
- WP9
- ISB
- ActiveNoiseMitigation
- Infrastructures
- Interferometer
- Calibration
- ModelDesignTools
- NoiseChar
- ObservatoryDesignAnd
- MaterialsDatabase
- Optics
- CoreOptics
- InputOutputOptics
- ScatteredLight
- SqueezedLight
- Wavefrontsensingand
- Suspensions
- LF_Payload
- Vacuum_Cryogenics
- Main

How to access MDC1 input data

The input data for the first ET Mock Data Challenge (MDC1) are distributed using the same infrastructure as the LIGO and Virgo collaborations use to distribute their data, namely the [Open Science Data Federation \(OSDF\)](#) (a.k.a. StashCache) plus publication in a [CernVM File System \(CVMFS\)](#) repository. The OSDF service is managed by [Open Science Grid \(OSG\)](#). The ET collaboration has its own Origin server within the OSDF, where the MDC1 input data files are stored. A network of OSDF caches ensures efficient access to those files from anywhere in the world. Publication in CVMFS facilitates access to the data by implementing a POSIX read-only file system.

This infrastructure offers different ways to access the data, as described below.

Access from CVMFS

The ideal, or simplest, situation corresponds to users running their analyses in computing clusters where the CVMFS client is installed and the CVMFS repository `et-gw.osgstorage.org` is mounted. In that case, the MDC1 input data files can be accessed on the following directory path: `/cvmfs/et-gw.osgstorage.org/et-gw/PUBLIC/MDC1/`.

For a fresh installation of the CVMFS client, we recommend to use the pre-configured CVMFS client package distributed by OSG. The corresponding documentation can be found [here](#). The CVMFS client can be installed on a user laptop. Files served by CVMFS can be copied from the corresponding `/cvmfs` path to a permanent location in a local hard-disk, effectively downloading the data. The purpose of CVMFS is not however to provide a mean to download data, but to facilitate (read-only) access to it.

Download from OSDF Origin with xrootd

The ET OSDF Origin server is in fact an XRootD server. Thus, one can use XRootD client commands to interact with the ET OSDF Origin. For instructions on how to install the XRootD client software, see the corresponding documentation from OSG [here](#). If you are working on a computer from an institution that is part of the [WLCG](#), XRootD client will most probably already be installed (you may need to source a setup script or load a module). Below are shown the XRootD client commands to:

- list the content of the `/et-gw/PUBLIC/MDC1` directory in the ET Origin server,

```
xrdls xroot://et-origin.cism.ucl.ac.be:1094/ ls /et-gw/PUBLIC/MDC1/
```
- recursively copy the `/et-gw/PUBLIC/MDC1` directory and its contents from the ET Origin server.

```
xrdcp -r xroot://et-origin.cism.ucl.ac.be:1094//et-gw/PUBLIC/MDC1 .
```

Download from HTTP web server

The MDC1 input data files can also be downloaded from an HTTP web server that is running on the ET OSDF Origin server: <http://et-origin.cism.ucl.ac.be> (see instructions on that web page).

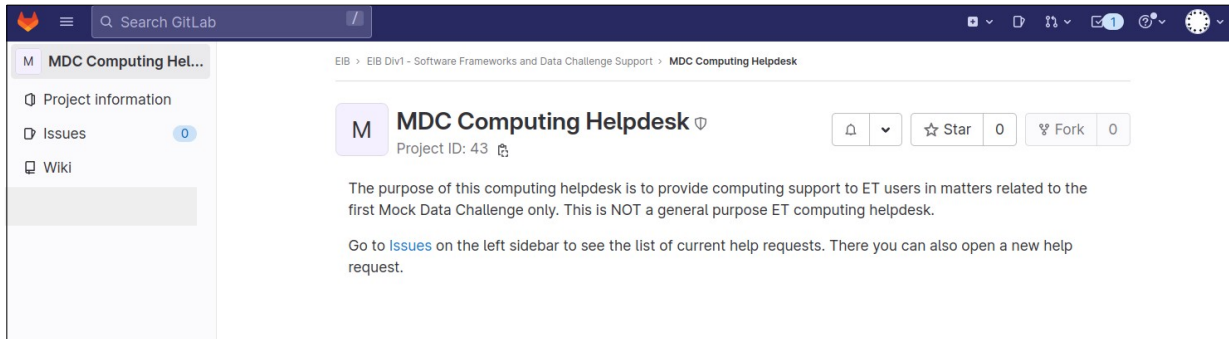
Computing support for MDC users

In case of problems accessing the MDC input data, or in general for any computing issue falling under the EIB responsibility and related to the MDC, you can open a ticket on the [MDC Computing Helpdesk](#).

MDC Computing Helpdesk

Created a “MDC Computing Helpdesk” project in ET GitLab under EIB/div1 group

<https://gitlab.et-gw.eu/eib/div1/mdc-computing-helpdesk>



Tickets via GitLab issues

Intended for MDC related issues under EIB scope

This is not a generic ET computing helpdesk

Summary

EIB Div 1 is providing support for ET Mock-Data Challenge by hosting the input data, distributing them to computing centres and offering means to download the data

- Created ET VO in OSDF

- Installed OSDF “origin” service at UCLouvain to host the MDC input data

- MDC input data from OSDF “origin” are published into CVMFS

 - `/cvmfs/et-gw.osgstorage.org/et-gw/PUBLIC/MDC1/`

 - Requires CVMFS client (available in WLCG and IGWN computing centres)

- Data are publicly accessible

- Download possible from UCLouvain’s OSDF “origin”

 - Requires XRootD client (available in WLCG and IGWN computing centres)

- Installed web server on UCLouvain’s OSDF “origin” to allow download without need to install any software

- Documentation in place in EIB Div 1 wiki

- Created MDC Computing Helpdesk (gitlab issues) for user support

Part 2

Proposed (short) list of tasks on which, I think,
we should start working on next

Task 1 – Software quality

Develop policies and best practices to ensure software quality, and encourage/enforce their adoption

This should and can be done from the very beginning

Possible starting point: Tania said that she would like to get her MDC data generation code rewritten in a more professional way

Not sure if she already found a person for this

Should we also enforce software quality to user pipelines, or only to official ET software?

Task 2 – Software distribution

Software will need to be distributed to computing centres, clouds and who knows where else

The most obvious candidate tool is CVMFS

I would like that someone takes the responsibility for managing software releases and publishing them to CVMFS (once we have some official software)

Need to install a CVMFS server

Would prefer an institute with long term commitment to this task

Task 3 – Data distribution

We are currently using OSDF + CVMFS infrastructure from OSG

Investigate other ways of distributing data that do not depend on a US managed infrastructure, but a EU one

One candidate is ESCAPE Data Lake

We will have a presentation tomorrow by Enrique Garcia Garcia, let's see what we learn

Task 4 – User's output data

We do not know what kind of output data/files the users are producing

It seems we do not need to provide any support for MDC users to store/present their analyses outputs

But will we have to do it in the future?

Task 5 – The task I like the most

Running analyses on distributed resources

Gather information on what is being used by other experiments with similar requirements

We should promote grid computing from the beginning to not run into the same issues LIGO and Virgo are facing with pipelines that can not run on the grid, because they rely on features that are not part of the grid infrastructure

Thank you!

Time for discussion...