Scenarios and ideas for data distribution in the next MDC

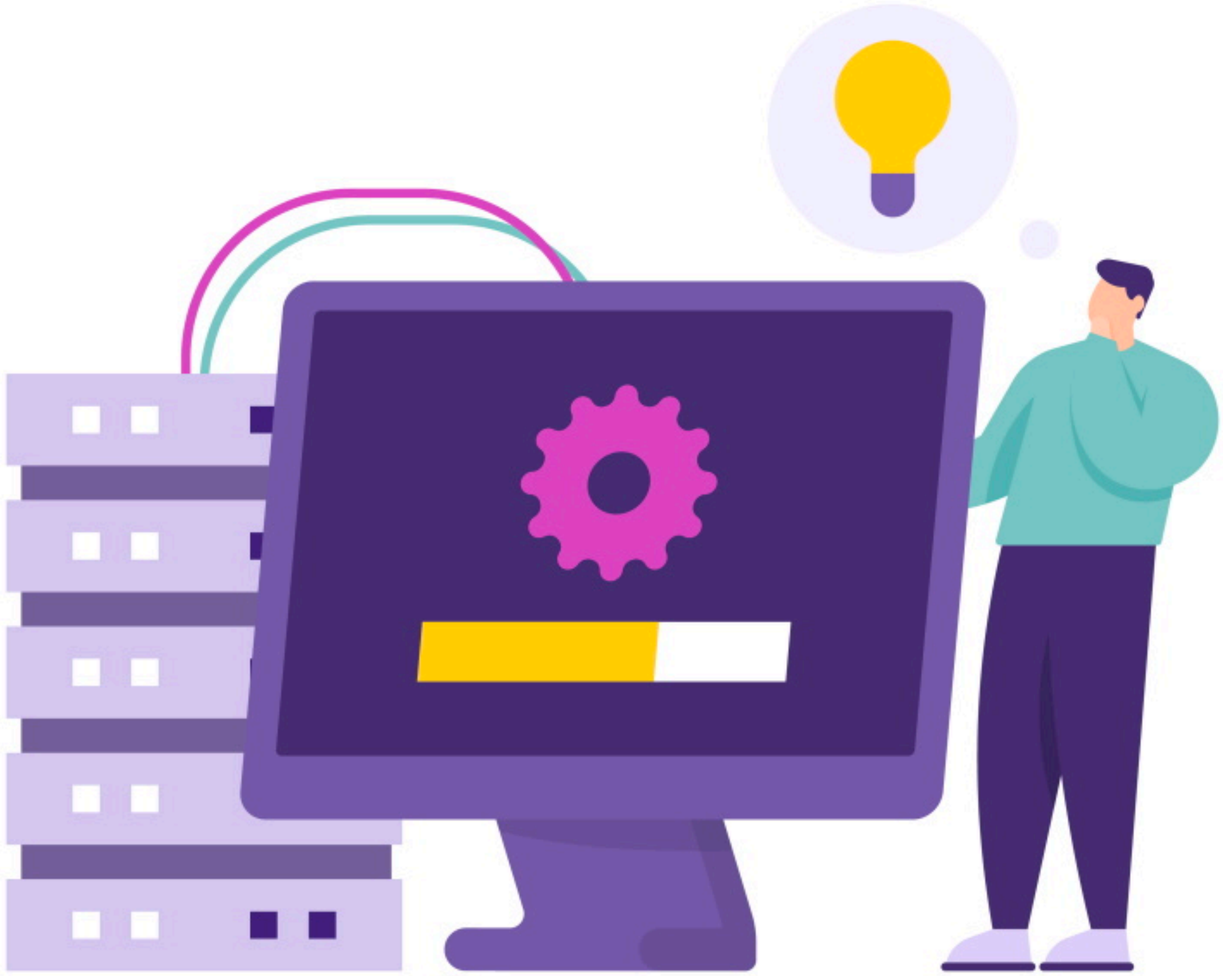
Einstein Telescope Monthly Meeting - October 4, 2022

Sara Vallero (INFN-Torino) for the ET e-Infrastructure Board (EIB)

Some considerations

Towards data delivery for ET

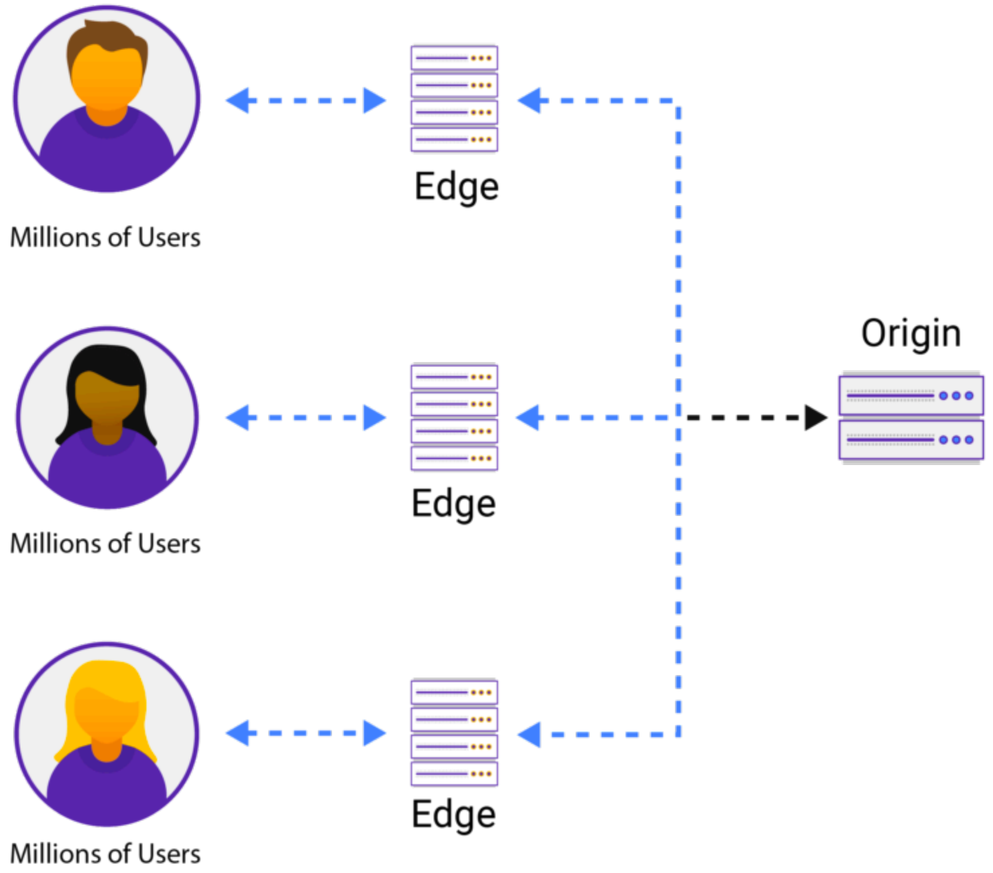
- We do not know the technologies that will be available in 10 years from now
- It will probably be a matter of keeping up to date with the evolving trends
- We should start early on to adopt promising state-of-the-art solutions



Content Delivery Networks (CDN)

The reason why your Netflix movies do not get stuck

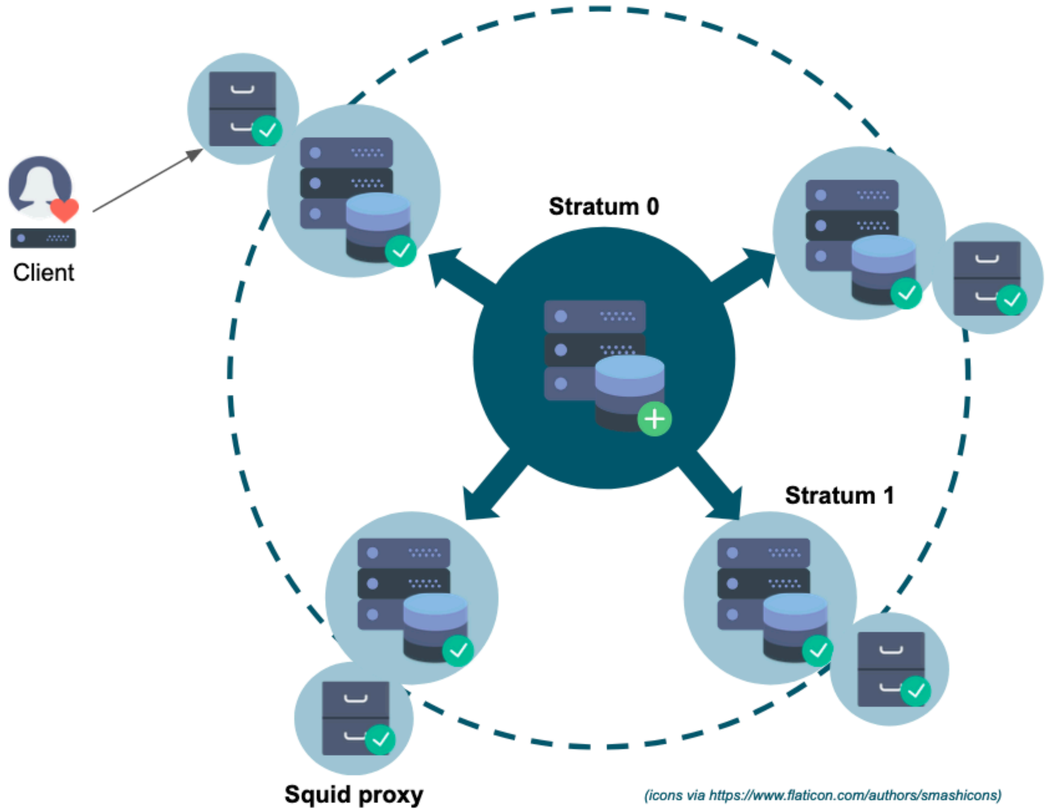
- Born in the late 1990s to avoid bottlenecks in the Web
- Geographically distributed network of proxy servers (edge)
- Provide high-availability and performance by placing the data close to end users



CernVM File System (CVMFS)

Developed to assist High Energy Physics (HEP) collaborations to deploy software on the WLCG.

- allows for efficient global distribution of software and data that does not change frequently
- caches files to disk so that, after the initial download, file access for the client is speedy
- implemented as a POSIX read-only file system in user space (a FUSE module)
- files and directories are hosted on standard web servers and mounted in the universal namespace /cvmfs
- can be used to distribute data and/or metadata



LVK = Ligo, Virgo, Kagra collaboration

CVMFS usage within LVK

https://computing.docs.ligo.org/guide/cvmfs/

- used to distribute both instrument data ("frame files") and analysis software for use at the shared computing centres and by distributed workflows
- Public repositories available: oasis.opensciencegrid.org, gwosc.osgstorage.org, singularity.opensciencegrid.org
- Private repository igwn.osgstorage.org:
  - access to the LVK proprietary data is restricted to registered LVK collaboration members
  - X.509 authentication

IGWN = International Gravitational-Wave Observatory Network (the LVK computing infrastructure)

The IGWN primary data distribution

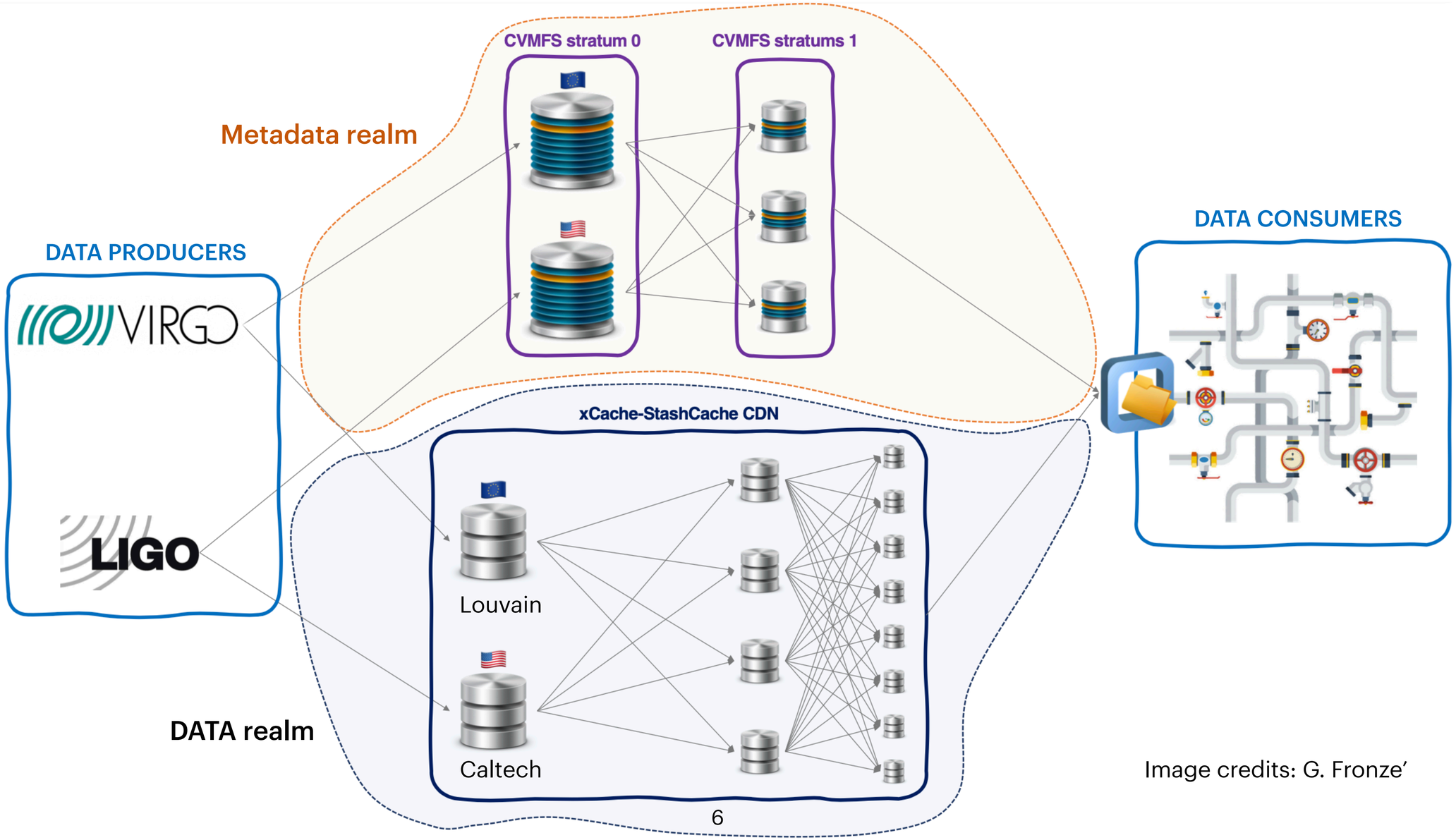


Image credits: G. Fronze'

Open questions
and possible scenarios

- Data size to be made available for the first ET MDC (~ 2 TB) ?
- How many users?
- How do we support people who are not in the LVK collaboration?

1st SCENARIO: leverage the IGWN CVMFS/CDN infrastructure

- number of users is not an issue
- but they all should belong to the LVK collaboration

2nd SCENARIO: setup an ad-hoc data distribution infrastructure

- fully fledged dedicated setup with CVMFS and cache CDN (might require long time to deploy, unless we can plugin in the existing OSG cache)
- minimal setup without CDN (data size and number of users might be an issue)
- where to host it?

3rd SCENARIO: make data available for download to non LVK members

- can be developed in parallel with one of the previous two

OTHER IDEAS?