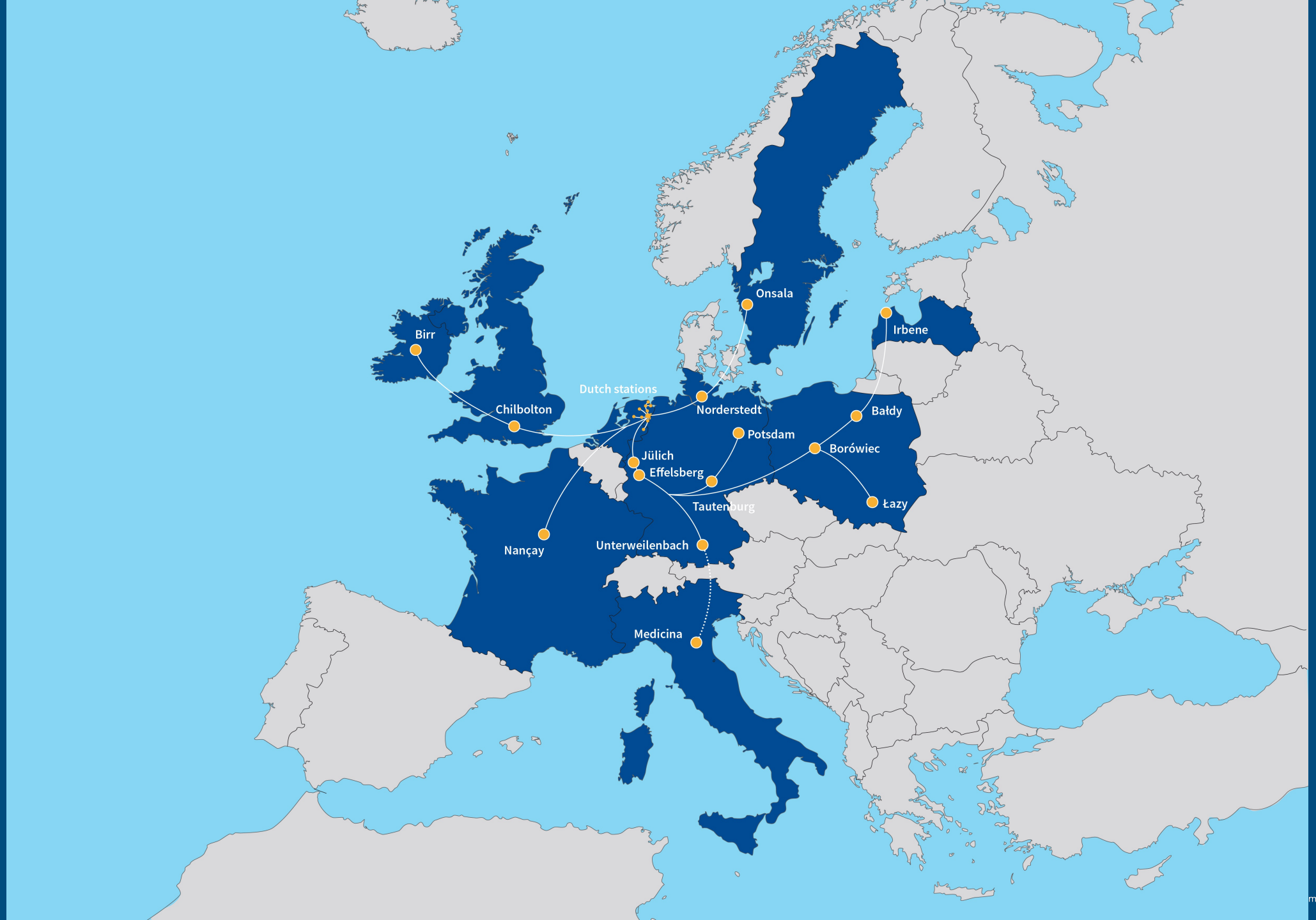

ASTRON
Netherlands Institute for Radio Astronomy

LOFAR data and processes

Yan Grange, Vishambhar Nath Pandey

ASTRON
Netherlands Institute for Radio Astronomy

LOFAR


Onsala
Irbene
Birr
Dutch stations
Chilbolton
Norderstedt
Bałdy
Potsdam
Borówiec
Jülich
Effelsberg
Łazy
Tautenburg
Nançay
Unterweilenbach
Medicina

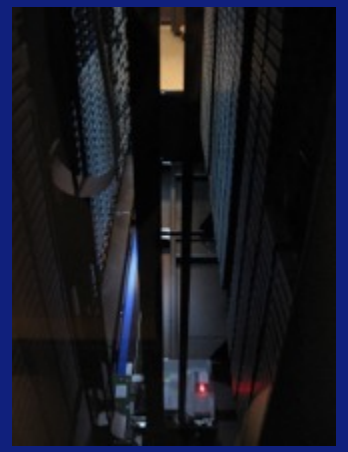
LOFAR data flow

Stations

Realtime system

Offline processing

Long Term Archive



240 Gbit/s

110 Gbit/s

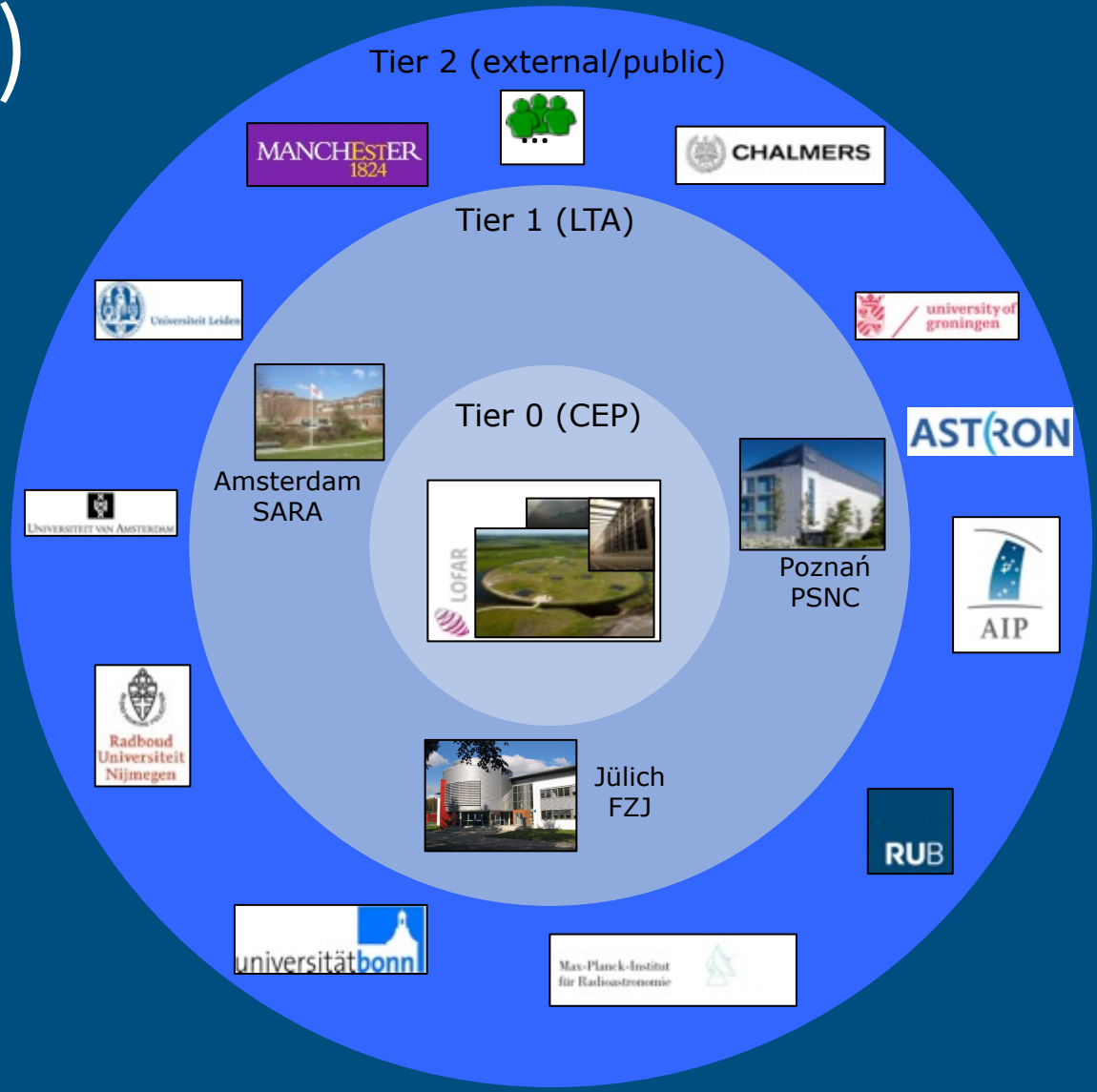
2 Gbit/s

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Long-Term Archive (LTA)

- 10 Gbit/s connectivity between Tier 0 and 1
- Networking infra shared with stations
  - Currently very little data movements within a tier.
- Rucio would map quite well here.
- All locations happen to use dCache.
  - We use gridftp for transfers
  - User access through http (either wrapped, or webdav with macaroons)



LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

LOFAR data flow

- Telescope generates data => pre-processing (**Groningen**)
- If processing is successful, data ingested in archive
  - If ingest is ok, data deleted (in some cases this may be need a manual action though).

- Central processing would be a Rucio Agnostic RSE? Or just a location with ingests?
- LOFAR LTA -> Read-only RSE (but with staging capacity). Is that Rucio Agnostic?
  - Anyhow current LTA locations as read-only RSE for DAC21 could be useful (but this would need to have a way to move data between VO's ☺)

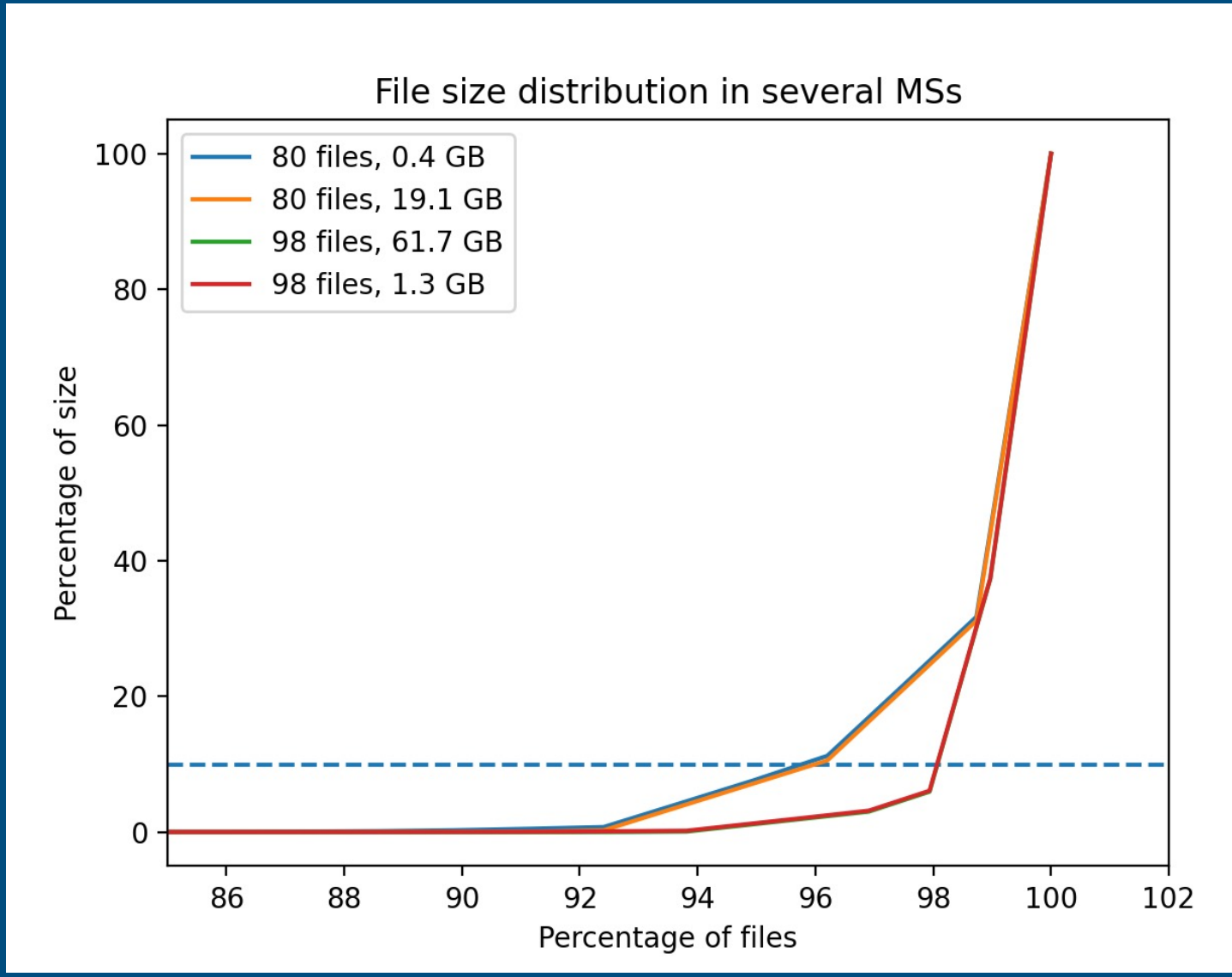
LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Data properties

- Instrument (lower level, until now)
- Higher level (target 2021-2023)
- Currently: Measurement set (MS)
  - Not a file, but a directory
  - In essence a database format. Main table is a list of antenna combinations and voltages for each time step.

- One observation (‘**dataset**’) consists of hundreds of measurement sets (‘**dataproducts**’). Each is a wavelength range.



LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Data organisation in the LTA

- Each granted observation proposal gets assigned to a **project**. Each observation gets and **obsID**. Each **dataproduct** in the observation will be assigned a **SubArrayPointing** and **SubBand** (wavelength band) number. Path in the archive then becomes:
- **project/obsID/**L**obsID**_SAP**SubArrayPointing**_SB**SubBand**_uv.MS_**hash**.tar
  (e.g. *lc0_012/152082/L152082_SAP000_SB138_uv.MS_243ca743.tar*)
  - Adding a hash to filename as poor mans AAI
  - Tarring files because of Tape storage
- Looks pretty much non-deterministic.
  - Non-deterministic RSE would make the structure visible in its own respect, seems more future proof.

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Data Life Cycle and AAI in general

- After data goes into the LTA, it is guaranteed on disk for a set period. Then disk copy is removed and only 1 copy on disk exists.
- For data access, data needs to be staged first (through our staging service.
  - Then again, data is pinned to disk for a set period.
- First year: only accessible for the PI (i.e. The person who wrote the proposal) has access. After the first year data is public (though staging and access based on basic authentication).
  - In principe the PI should be able to decide who has access to their data.

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

ActivityOne

- Current LTA -> read-only RSE -> is that what we want to call rucio-agnostic? Only thing that counts is that we can have data in there that we can read.
- Data transfers using current tooling, registering only later using Rucio (should be easy with non-deterministic RSE, but not impossible with deterministic either). *(Looks quite similar to the MAGIC use case).*
- We already have our own dir structure, advantage of non-deterministic RSE is that it makes our main data location to behave like we expect. Disadvantages?
- Data life cycle does strongly connect to QoS and AAI.

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy