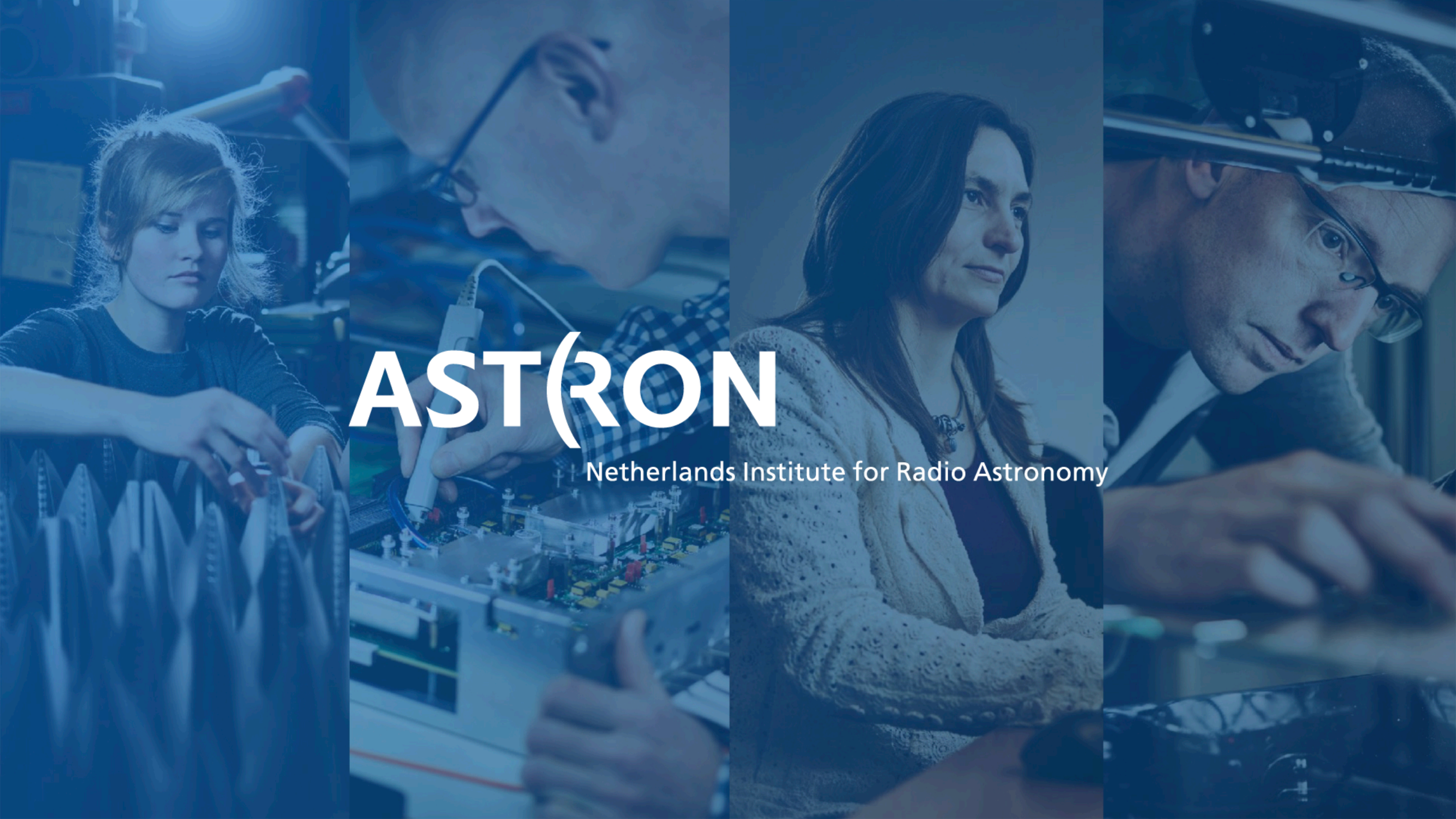

ASTRON
Netherlands Institute for Radio Astronomy

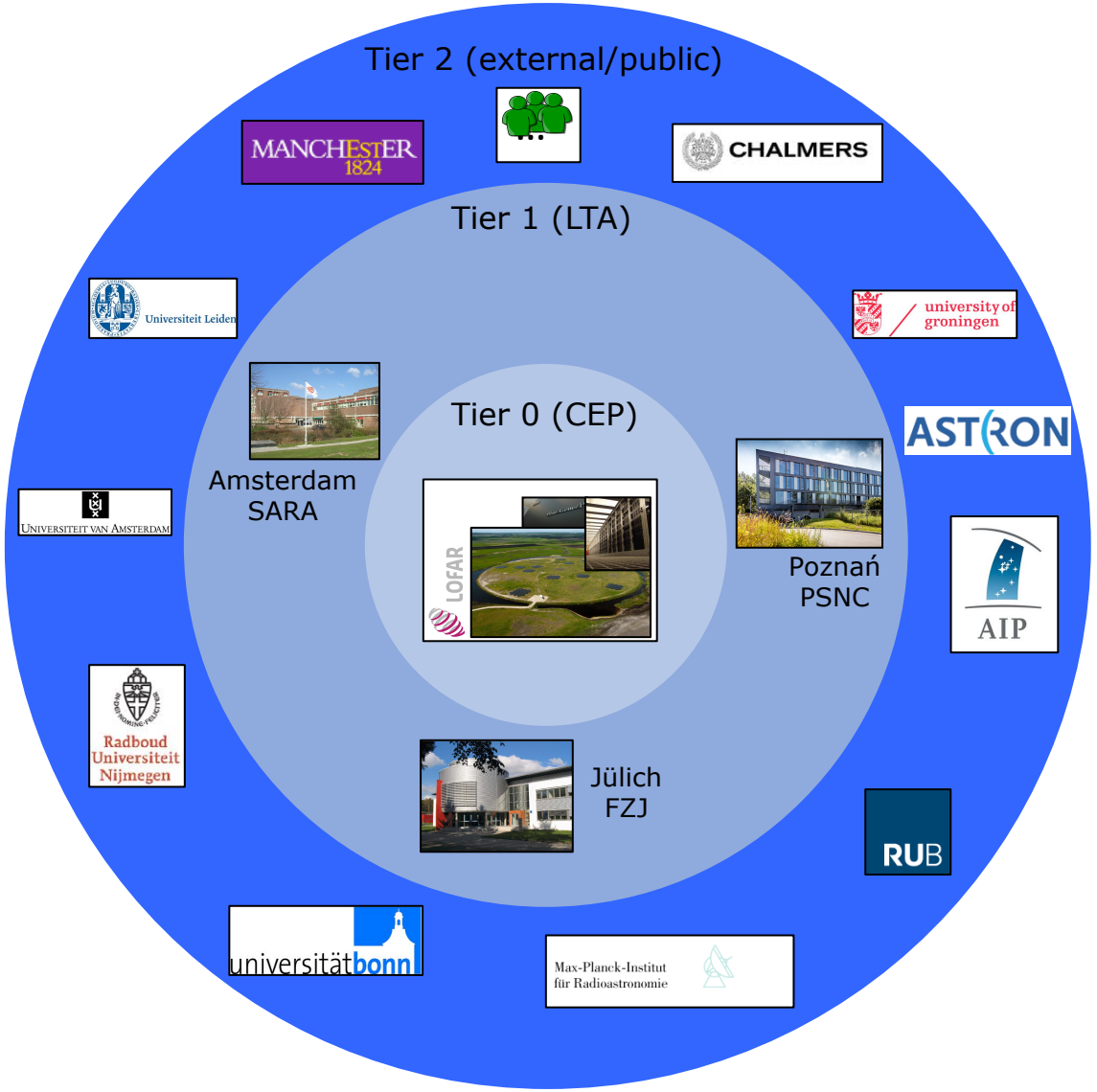
LOFAR Long-Term Archive data transfers

Yan Grange, Hanno Holties, Jorrit Schaap, Adriaan Renting

ASTRON
Netherlands Institute for Radio Astronomy

LTA architecture

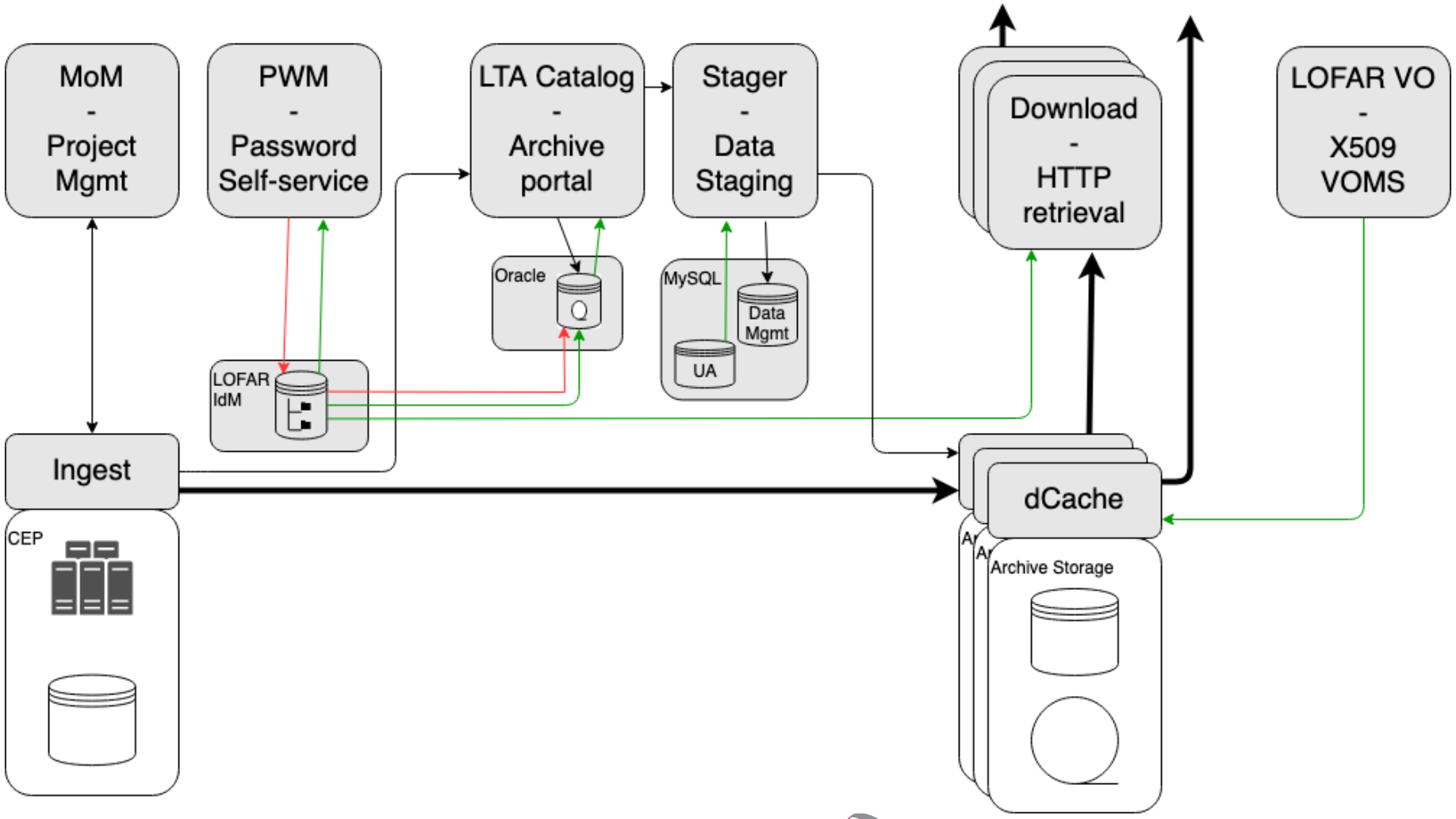
- Observations are pre-processed in Groningen (Tier 0)
- Single copy on one of the three archive sites
- Life cycle:
  - Data copied to disk pool where it is guaranteed for one week
  - Long-term storage on tape only



LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

LTA architecture

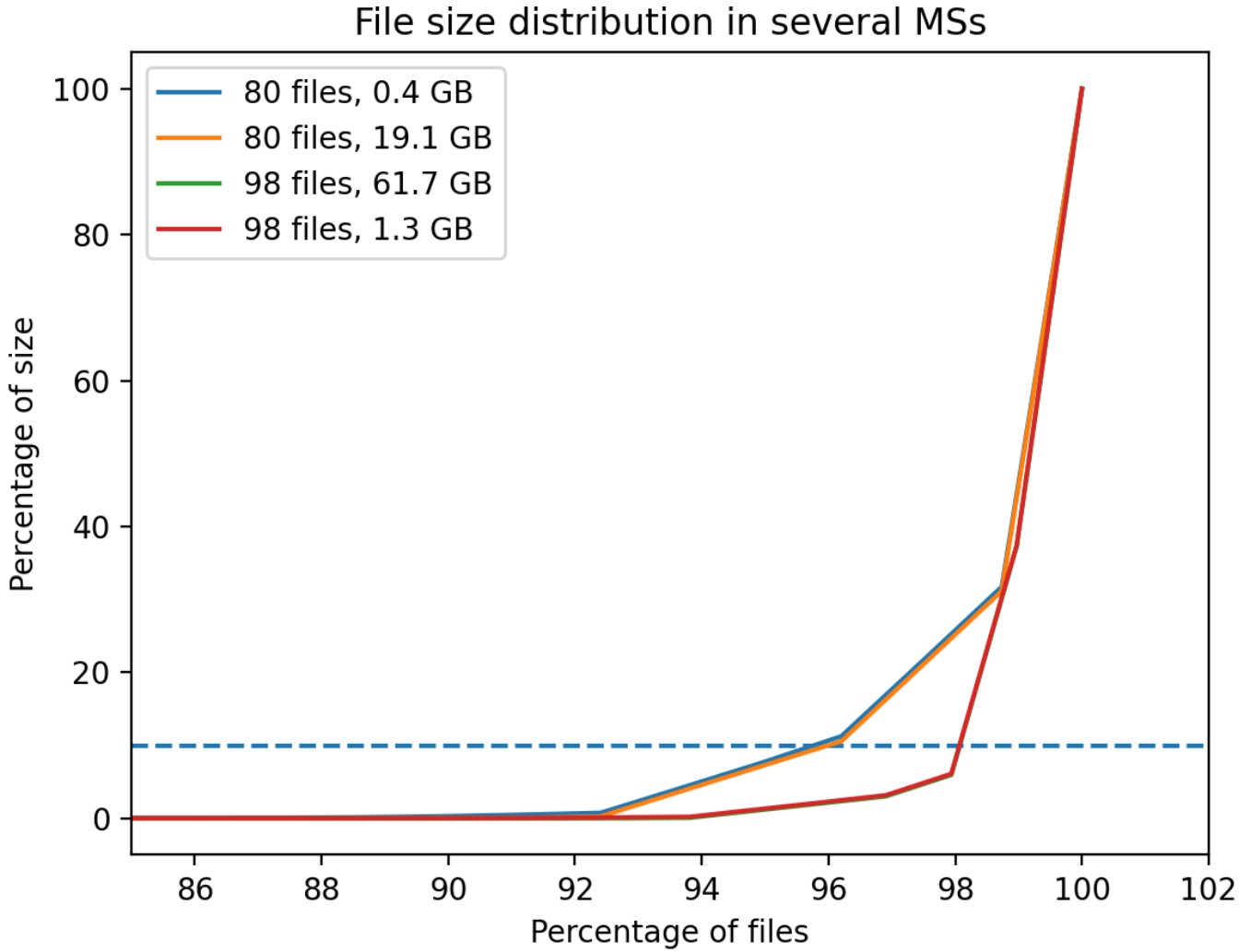


LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Data properties

- Instrument (lower level, until now)
- Higher level (target 2021-2023)
- Measurement set (MS)
  - In essence a database format. The content is a list of antenna combinations and voltages for each time step.
  - Directory structre
  - Typically ~100 files with ~1 containing the bulk of data
- One observation (‘data set’) consists of hundreds of measurement sets (‘data products’)

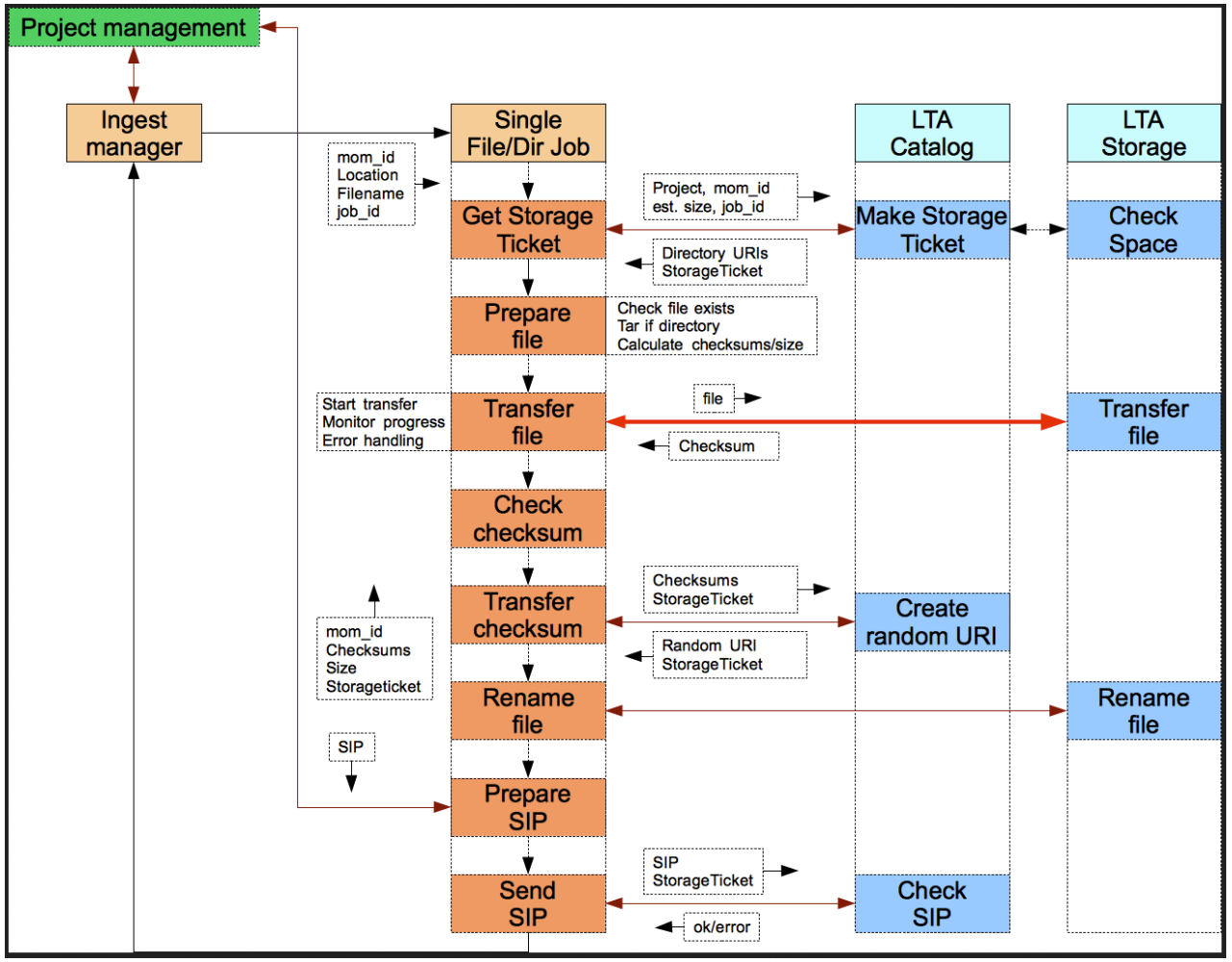


LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Ingest procedure

1. T0 obtain storage ticket from T1
2. T0 prepare data products
3. T0 send data products
4. T0 retrieves checksum from T1
5. T0 verifies checksums of data provided by T1
6. T0 puts checksums in LTA catalog and obtains final path
7. T0 renames URL to final path
8. T0 prepares and submits metadata to LTA catalog



LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Data Transfer - Considerations

- Minimise disk & network IO
- Minimise disk capacity (avoid replication)
- Verify data integrity during transport
- Allow user verification of integrity after retrieval from LTA (i.e. MD5)
- Package data products in tape friendly manner (i.p. MS, but also e.g. output PULP))
- Ability to scale out (multiple threads & servers)
- Embed in transaction type of data ingest process

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Existing solutions

- Transfer tools (GridFTP, HTTP, but also iRODS)
  - Performance through ‘network bashing’ (divide files in chunks, send in parallel)
  - Multiple reads (checksum, transfer, checksum)
  - Sequential handling of individual files
- At time of development no easily adoptable data management tools (a la Rucio) available
  - Available ones not trivial to integrate with piped/streaming packaging & checksum verification.
- Also rather specific use case, relaxed usability requirements (i.e. no ‘non-expert’ users).

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

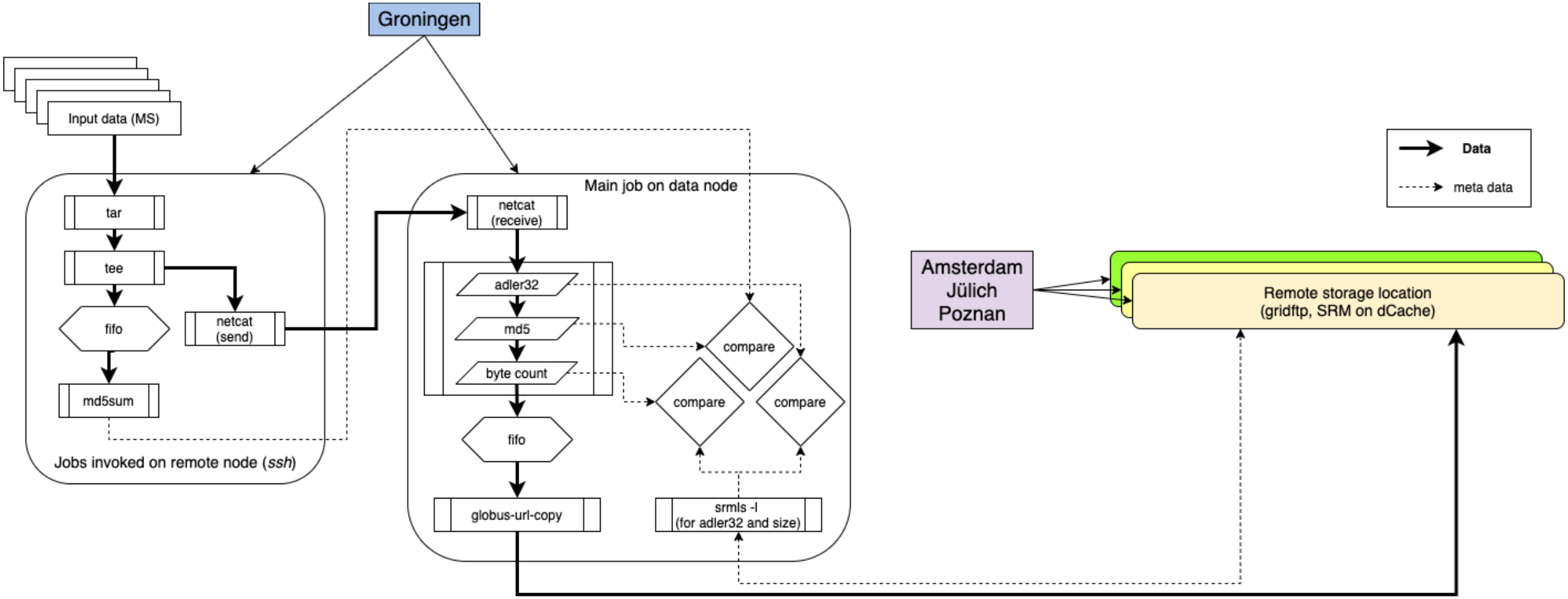
LtaCp

- SRM was originally the default for all data transfers, and still is the main retrieval method (work in progress on using webdav).
  - Gridftp blocks while remote site is computing checksums. This would cause timeouts in SRM in the past. We did move to direct use of globus-url-copy early on, mainly for its capability to read from pipes
- Original LtaCp written in Java. Current implementation uses python and default Linux tools (netcat, tee, tar) plus custom streaming md5+adler32 calculation and byte count.
- For long-distance transfers, globus-url-copy (with robot certificate) is used:
  - Streaming data through (see next slide)
  - No data channel authentication (our data is not that sensitive) [-nodcau]
  - 4 parallel threads (unclear if this actually does anything when streaming) [-p4]
  - Create directories [-cd]
  - Buffer size 131072 [-bs 131072]
- For performance we have 20-40 concurrent data threads. Close to filling 10Gbps, even when some transfers wait for remote checksum computation to finish

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

LtaCp – current implementation



LOFAR

ASTRON
Netherlands Institute for Radio Astronomy

Lessons Learned

- JAVA vs Linux
  - Flexibility
  - ‘Just retry’ vs interpreting (volatile) error codes/messages from the transfer tool
- Timeouts: If no traffic over connection session, active network components along the connection may drop the session (receiving servers, routers, firewalls). In particular on control connection during transfer and during checksum verification.
- Getting jumbo frames consistent along the line can be (often is) a pain (multiple organizations, first pointing at others; intermediate components not responding to ping, complicating troubleshooting).
- Same for misbehaving network equipment. Typical first response: ‘works for me’. Can take days/weeks of persistent nagging, convincing, and reverse engineering of the network to home in on culprit. PerfSonar can help once the network has been proven to be good.

LOFAR

ASTRON
Netherlands Institute for Radio Astronomy