Advanced Data Staging in the ARC Computing Element

David Cameron

University of Oslo

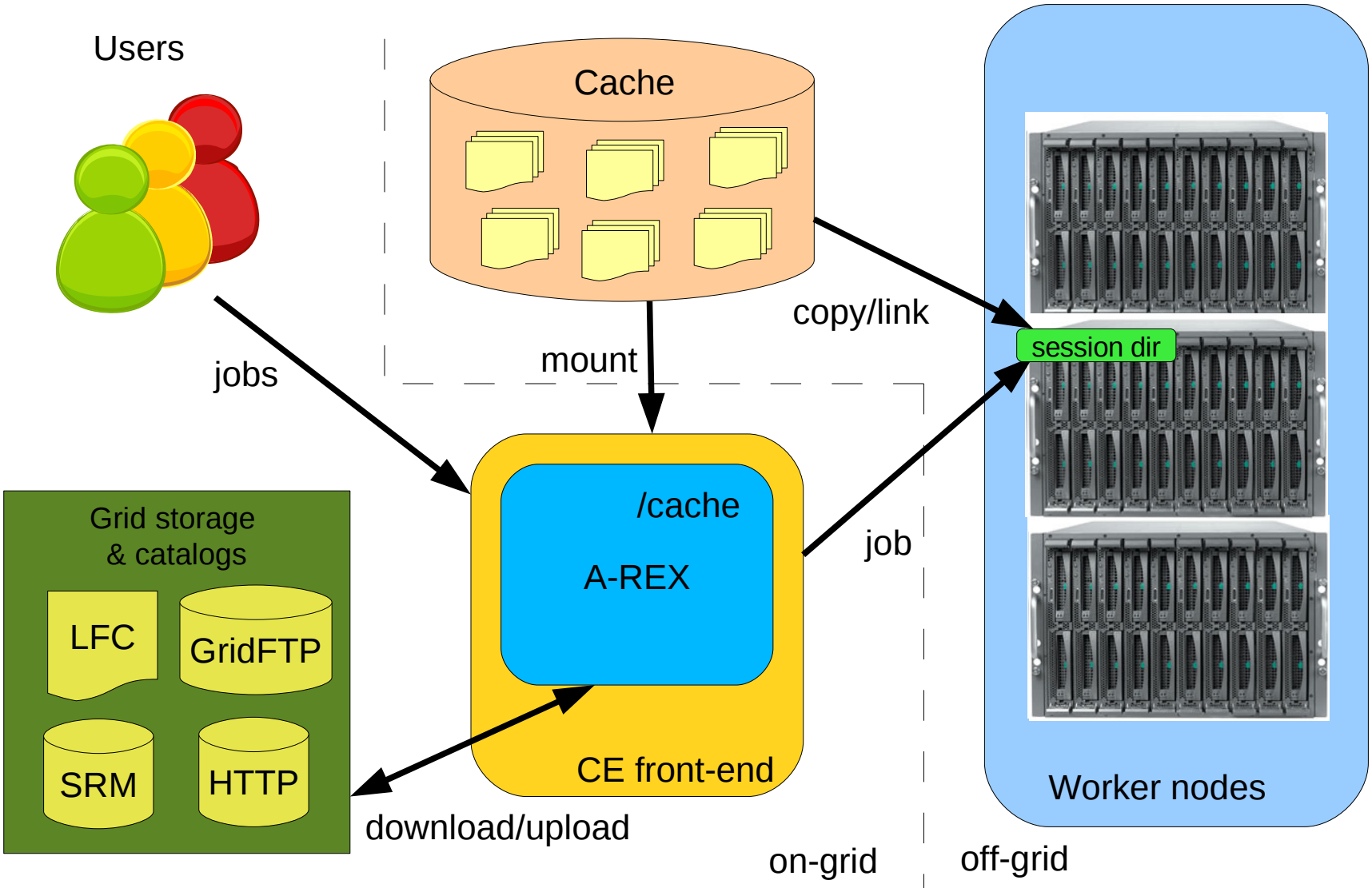
- EGI Community Forum, Munich, 27.3.12

Acknowledgements

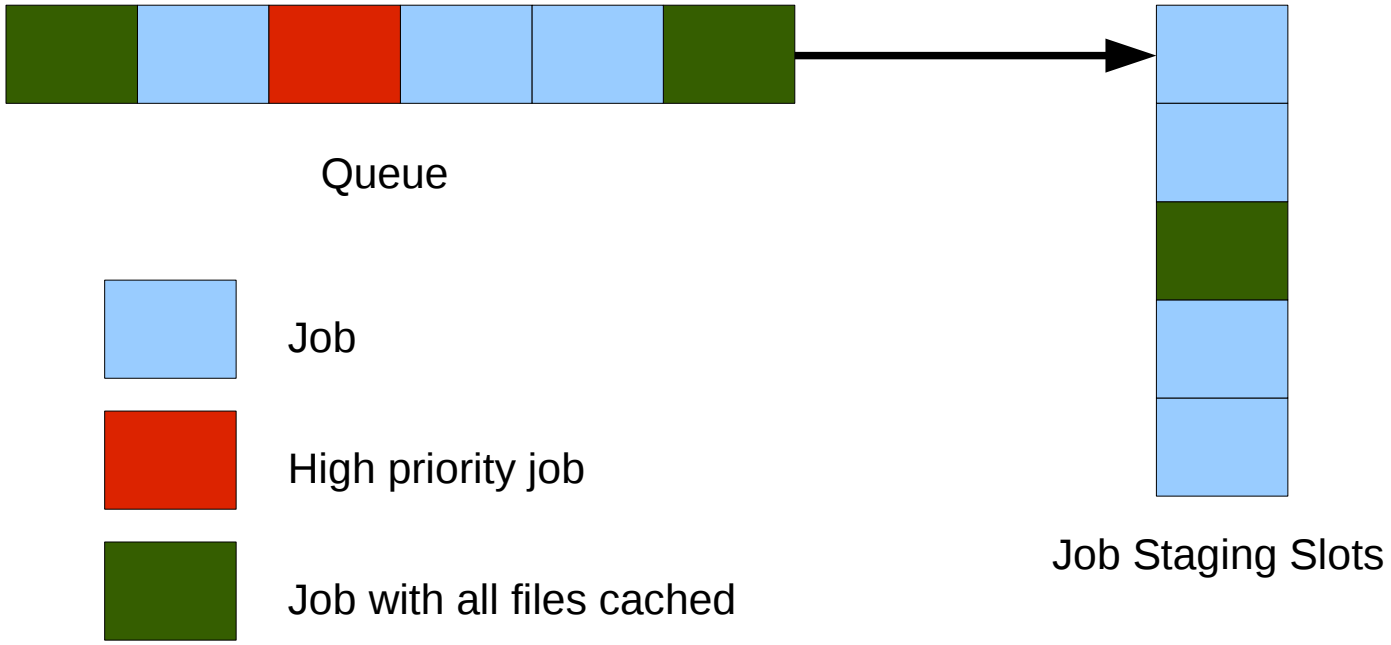
- Co-designers/developers:
  - Aleksandr Konstantinov (Univ. Oslo)
  - Dmytro Karpenko (Univ. Oslo)

- Early adopter
  - Andrej Filipic (IJS, Slovenia)

ARC Data Management Architecture

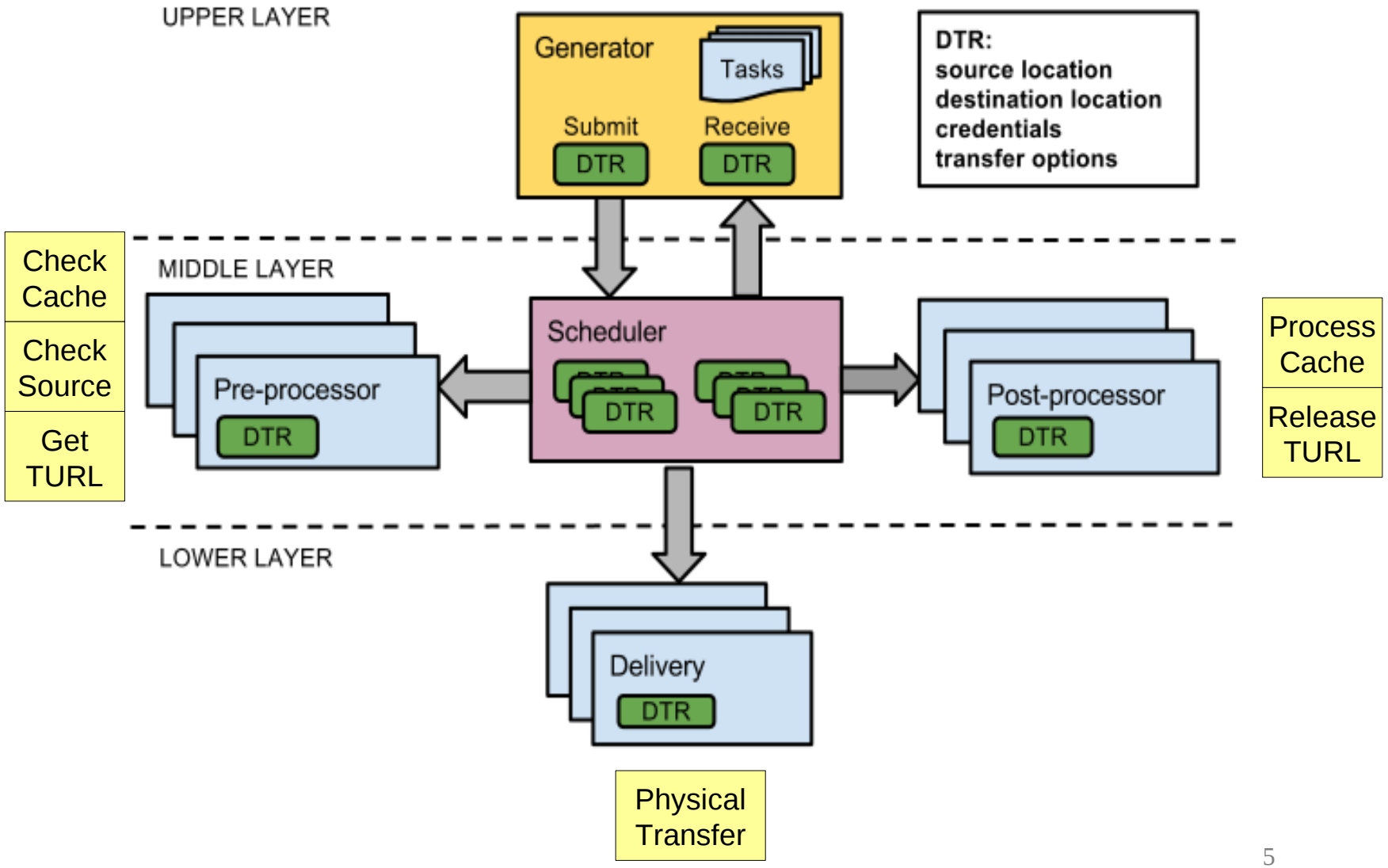


Previous system



- FIFO
- No priority system
- SRM wait calls are blocking
- Unit of transfer is job
- Cached files are blocked by transfers

New Design: Data Transfer Request



Advantages

- Site-wide queue of all files
- Independent queues for pre- and post-processing
- Instant cache processing
- Asynchronous SRM staging calls
- Priority system
- Can swap one layer for another product

Implementation

- Lower two layers are new library libarcdatastaging
  - Offers API for generic application
- Generator is implemented in A-REX
- Each component runs as separate thread
- All state in memory
  - Extremely fast queue processing
  - No external dependencies or components (eg DB)
  - Summary of state is frequently dumped to file for monitoring/recovery
- Bulk operations for LFC and SRM in pre-processor
- Intelligent error handling

Supported Protocols

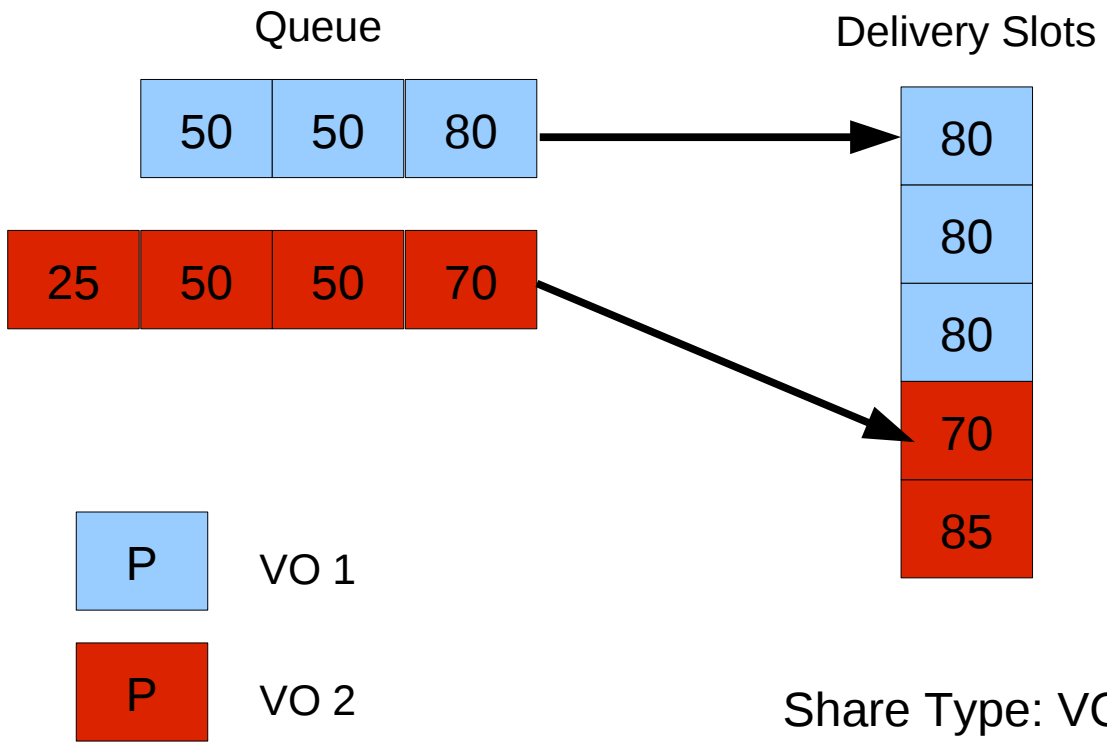
- File
- HTTP(s/g)
- GridFTP
- SRM
- LFC
- Xrootd (read-only)
- GFAL (experimental)
- RLS (deprecated)

- A DTR can be a transfer between any two protocols, but no 3rd party transfer

Shares and Priorities

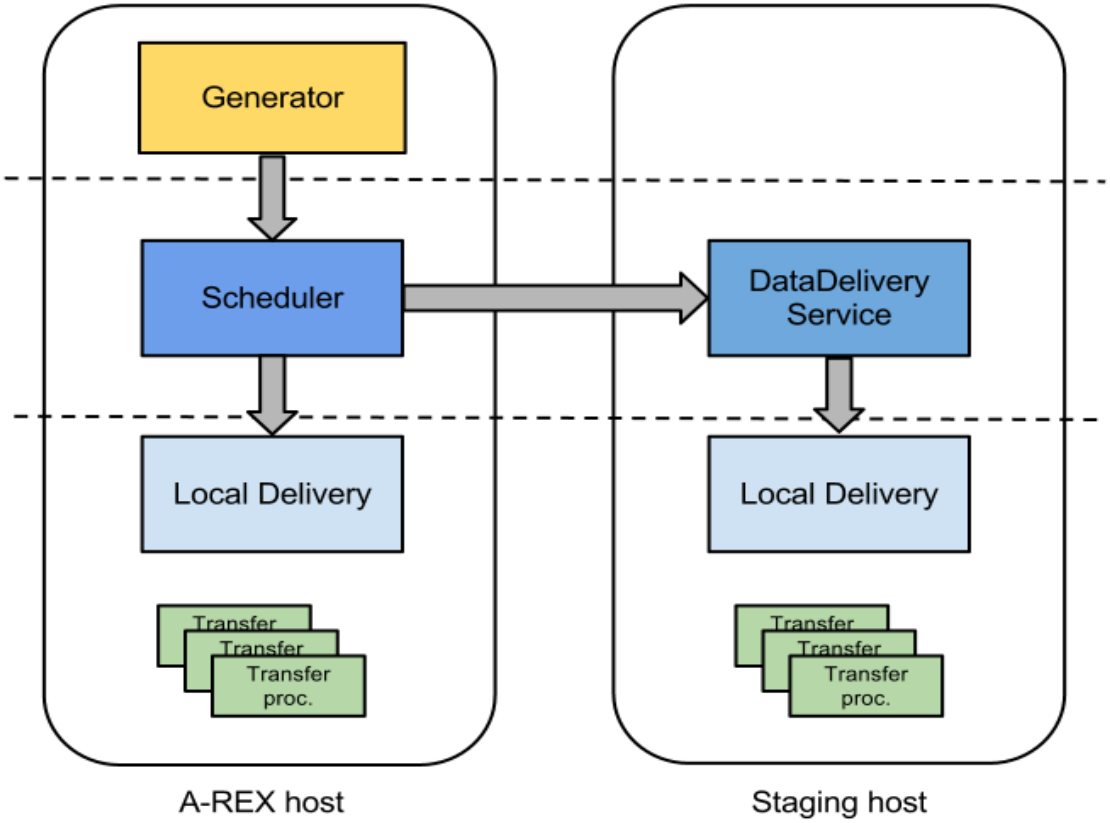
- "Shares" are a way to group DTRs by DN, VO, VO role etc.
  - Slots are split among shares to avoid one share blocking others
  - Share priorities decide the number of slots per share
- Priorities can be defined (integer 1-100)
  - On the CE for certain shares
  - In the job description by users
- This means sites can determine share priorities but users determine priority within their share

Shares and Priorities



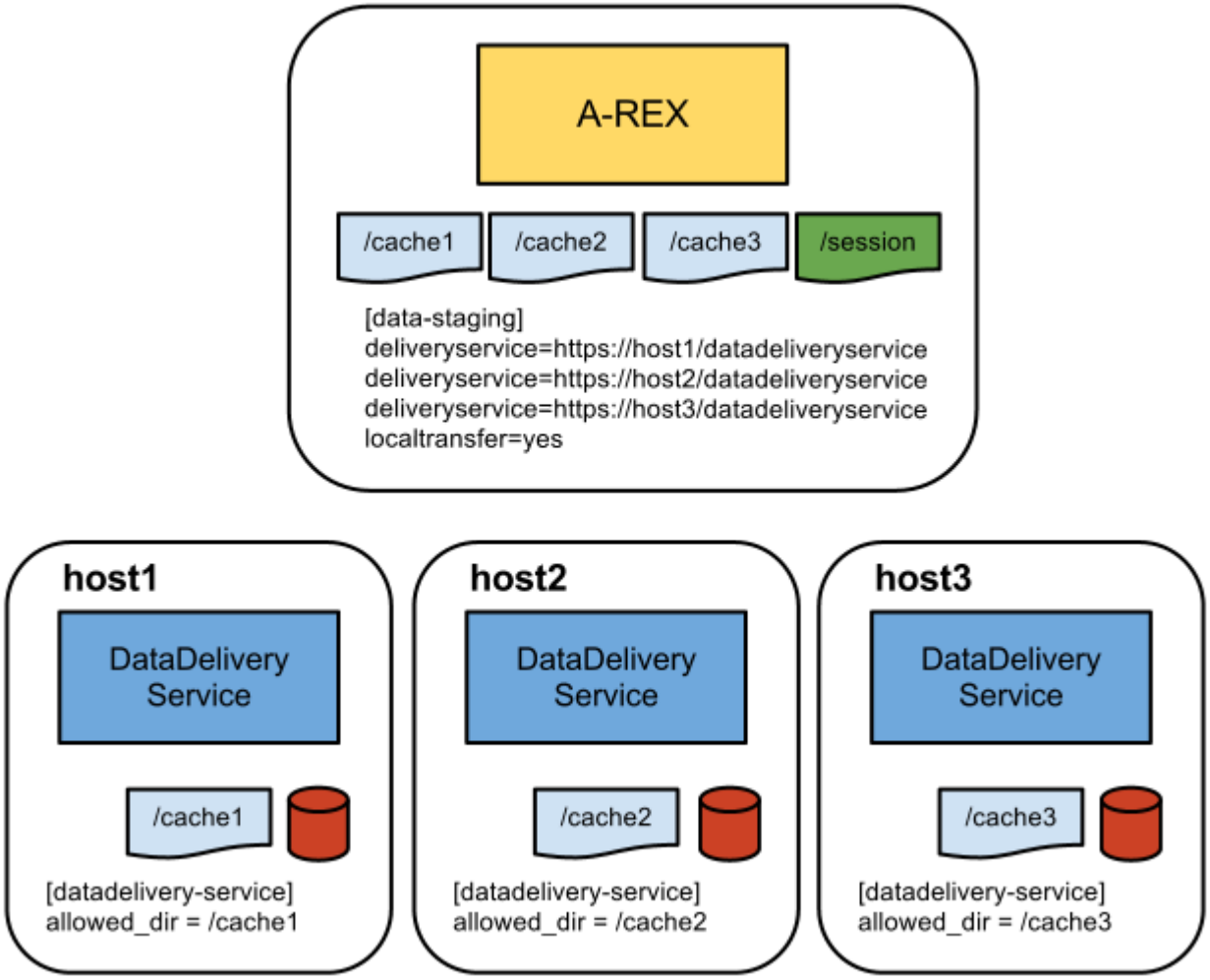
Share Type: VO
Priority VO 1: 60
Priority VO 2: 40
Max transfer slots: 5

Multi-host Staging



- Delivery layer can be expanded to multiple hosts
- Data delivery service allows Scheduler to launch transfers on remote hosts

Deployment Scenario



- More info in tutorial on Thursday afternoon

Monitoring

- DTR state info is periodically dumped to a file
  - Easily parse-able by scripts

```
c547fe20-7f92-4496-bdf9-33347114e608 STAGING_PREPARING 25 atlas:null-download
0c0a6ddb-12d0-4c46-81a9-da5852a6bb27 STAGING_PREPARING 25 atlas:null-download
8bd344aa-6164-46df-94bc-00cf267ae7dd STAGING_PREPARING 25 atlas:null-download
73fab80f-5410-4e1a-a554-33c87397662b CACHE_PROCESSED 25 atlas:null-download
c13eed22-ec3a-4893-a8ef-d87575514ba1 CACHE_PROCESSED 25 atlas:null-download
95664907-6947-4449-863e-d323c66bd48c CACHE_PROCESSED 25 atlas:null-download
359150d0-afcf-49da-9d4a-dea3ac94b4cd PROCESSING_CACHE 25 atlas:null-download
852c94ee-6254-4d29-8e22-262415eb824f PROCESSING_CACHE 25 atlas:null-download
```

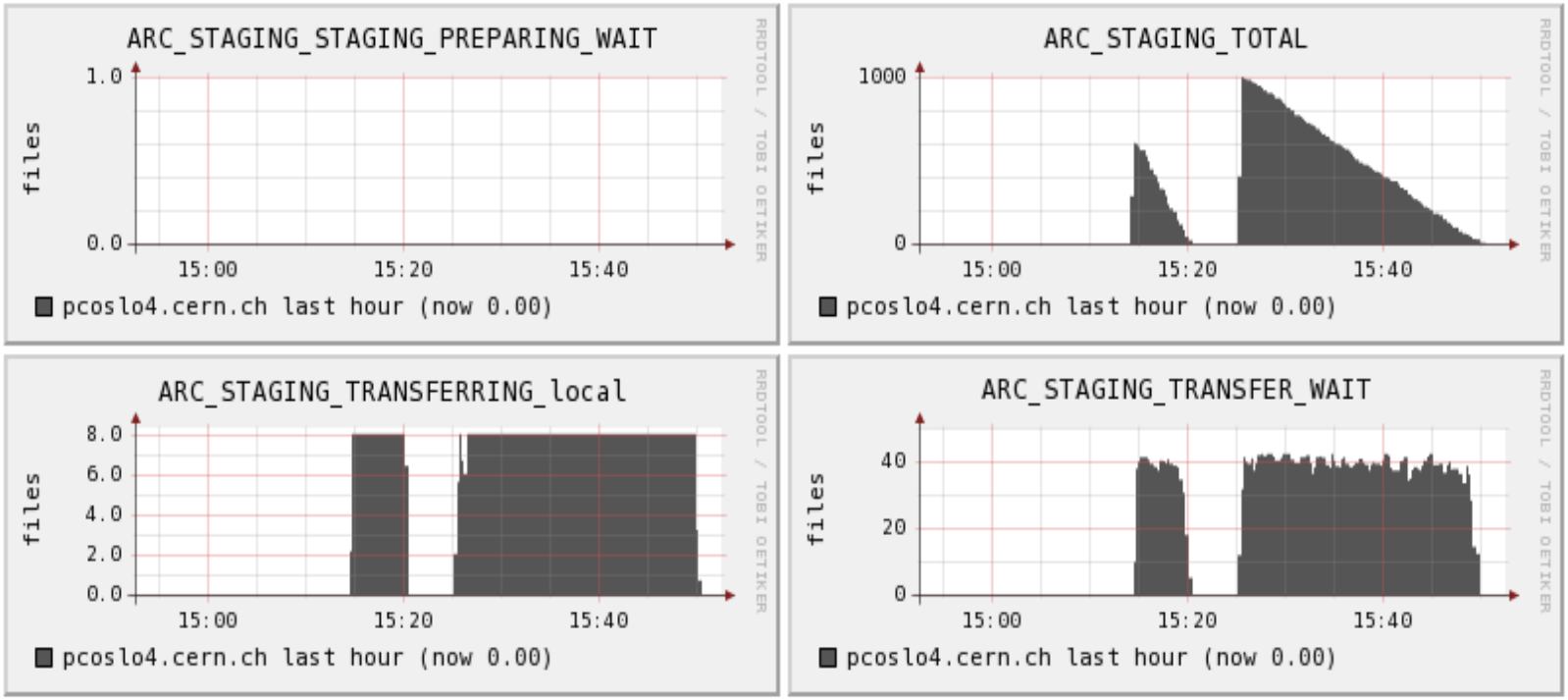
```
> transferstates.py
RESOLVE 232
QUERY_REPLICA 1
TRANSFER 42
TRANSFERRING 8
RESOLVING 106
PRE_CLEANED 312
CACHE_CHECKED 10
QUERYING_REPLICA 30
CHECK_CACHE 243
```

```
> gm-jobs -s

Preparing/Pending files        Transfer share
        8/536                  atlas:null-download
```

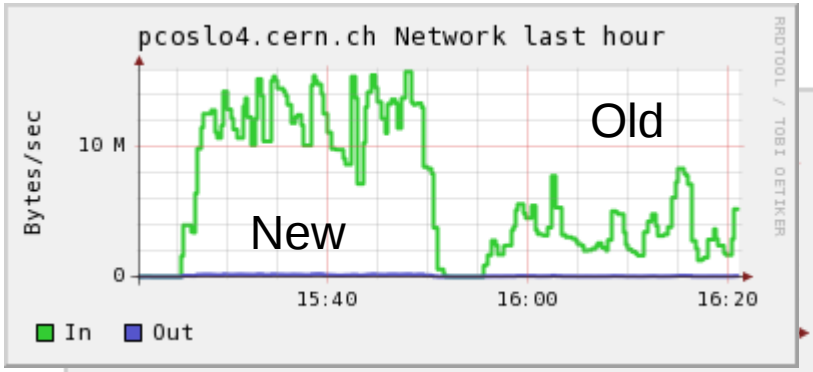
Monitoring

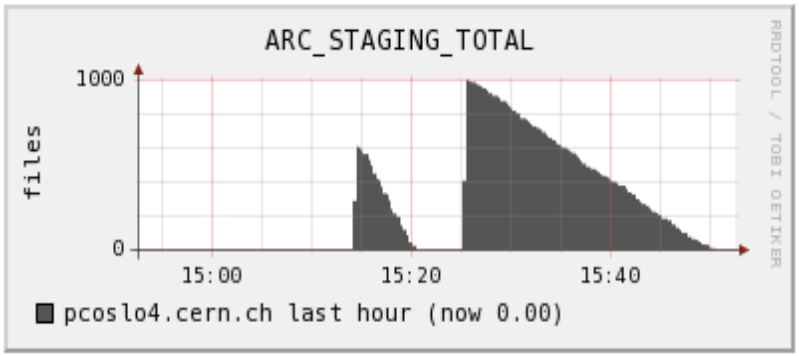
- Graphical monitoring via Gangliarc
  - Adds ARC-related ganglia metrics via gmetric

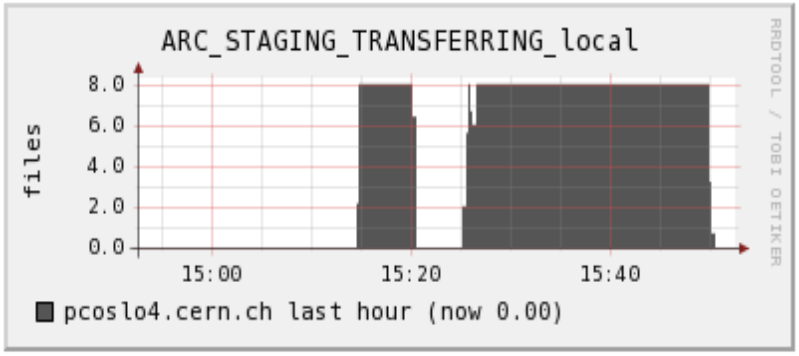


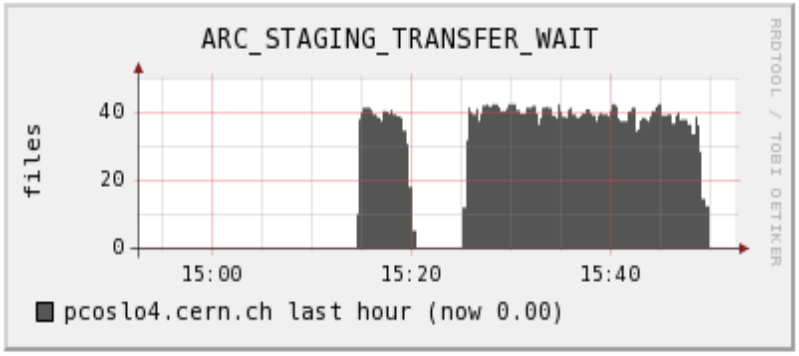
Results

- Much more efficient use of network resources
  - A new transfer is always ready to start
  - Bulk operations



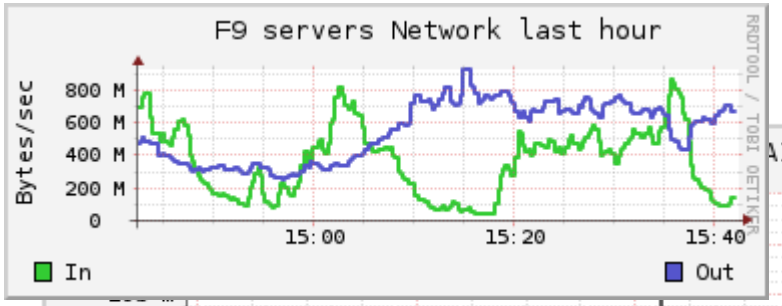


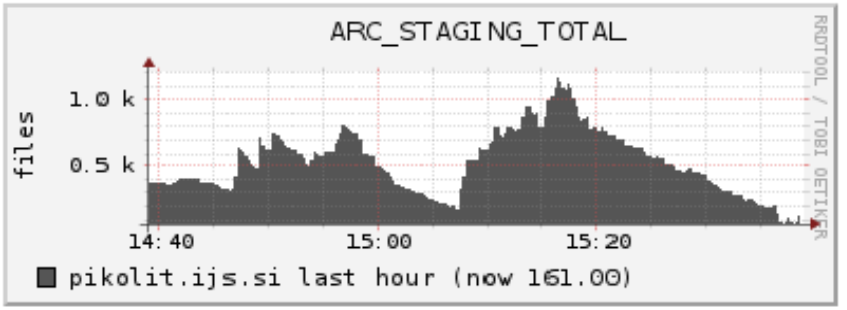


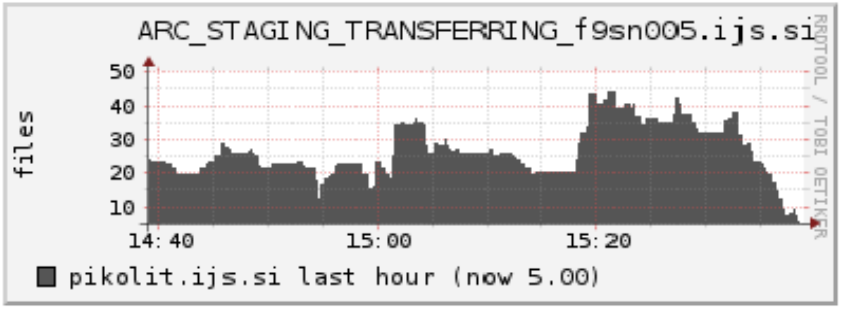


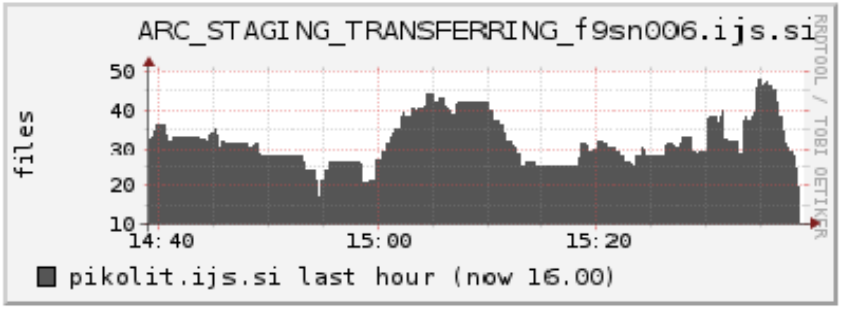
Real Life Usage

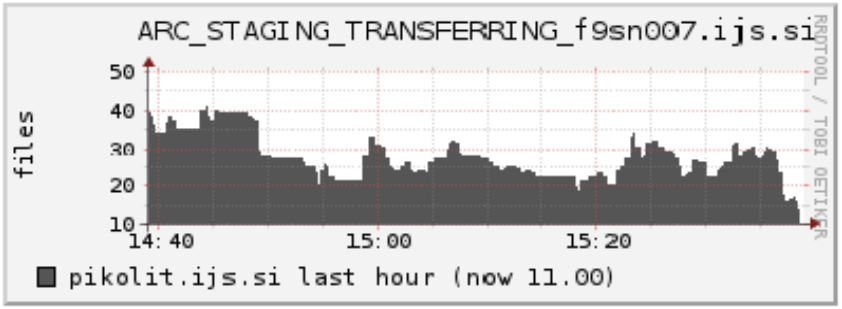
ATLAS production/analysis at pikolit.ijs.si

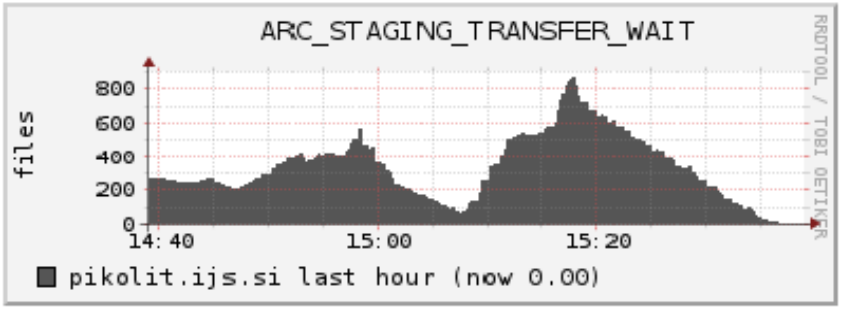












Future Plans

- First fully-functional release in EMI-2 ARC CE
  - newdatastaging=yes to enable
- Collect feedback from sites and make improvements
- Improved monitoring for end-users
  - To know transfer states of their jobs
- More intelligent priority algorithms
- Java and Python bindings for library

Links

- Main wiki page
  - http://wiki.nordugrid.org/index.php/Data_Staging
- Multi-host data staging
  - http://wiki.nordugrid.org/index.php/Data_Staging/Multi-host
- Gangliarc
  - http://wiki.nordugrid.org/index.php/Gangliarc
- Tutorial on Thursday
  - https://www.egi.eu/indico/sessionDisplay.py?sessionId=45&confId=6

Thank you!

EMI is partially funded by the European Commission under Grant Agreement RI-261611