HIFIS backbone transfer service: FTS for everyone

Tim Wetzel¹, Paul Millar¹, Uwe Jandt¹, Patrick Fuhrmann¹, Stefan Helmert², Tobias Huste², Guido Juckeland²

¹ Deutsches Elektronen-Synchrotron DESY, ² Helmholtz-Zentrum Dresden-Rossendorf HZDR

EGI conference, Nov 03, 2020
Helmholtz Association (HGF)

- Founded in 1995 to formalise relationships between research centres
- Members: 19 autonomous research centres in Germany
- Mission: Contributions to grand challenges facing society, science and industry
- Fields: energy, earth & environment, health, aeronautics, space & transport, matter and key technologies
Motivation

New challenges

- Top position of Helmholtz research relies increasingly on **cross-centre** collaboration and international cooperation
- Growing importance of cloud access to **common data treasure** and -services
- Rapidly growing **data exchange** from research instruments requires excellent data networking
- Growing connections between HGF, **EOSC** and **FAIR**
Helmholtz Incubator

- Helmholtz aims for joint research & information environment for all Research Fields
Helmholtz aims for joint research & information environment for all Research Fields

- Helmholtz Federated IT Services HIFIS
- connects all HGF centres
  - ....and their world-wide collaboration partners!
- Portfolio of federated IT services
- Common marketplace
- Access via HelmholtzAAI with home credentials

Easy and comfortable federation for collaborative web services

Photos by Markus Winkler, NOAA, National Cancer Institute on Unsplash
Helmholtz aims for joint research & information environment for all Research Fields

- High performance + collaborative services
- Shall connect all centres
- ....and their world-wide collaboration partners!
- Secure, simple access and easy-to-use

Widely establish best-practices for development+use of research software:

- high level of quality, visibility and sustainability

This year’s pandemic was a powerful reminder of how important collaborative and scalable IT services are today!

Photos by Markus Winkler, NOAA, National Cancer Institute on Unsplash
Why data transfers?

- Helmholtz centres **distributed** all over Germany
- Large data sets in **collaborative** research projects
- **Policy-driven** data transfers required
- Data analysis often sensitive to **latency**
- **Data locality** is important!
- Part of **HIFIS** backbone contract
  - Reliable, comfortable and robust transfer methods needed
Transfer services

- As HIFIS backbone core service
- CERN’s FTS3 as backend
- webFTS as comfortable WebUI
- FTS3-REST-API as CLI for automated transfers
- Later: Rucio for policy driven transfers

→ Client applications for all needs and purposes
HTTP Third-Party Copy (TPC)

- WLCG development for data transfers
  - Extension of the HTTP protocol
  - Third party can commission transfers between source and destination
  - Data is transferred directly between endpoints w/o third party
  - One endpoint needs to understand TPC-COPY extension (active party, WLCG)
  - The other endpoint needs to enable PUT or GET requests for files (passive party)
- Asynchronous data transfers possible
  - Not implemented in standard Apache httpd
  - dCache needed as active party
WLCG storage Endpoints

- storage solutions in WLCG:
  - dCache
  - EOS
  - DPM
  - StoRM
- Developed for constant high load and huge data volumes
- Enclosed view on data
- More open endpoint solution needed for HIFIS
Endpoint components

- Apache **httpd** webserver modules used:
  - mod_ssl (SSL/TLS capabilities)
  - mod_dav (webDAV capabilities)
  - mod_auth_openidc (OpenIDConnect/OAuth2)
  - modified mpm_itk (Multiprocessing module, user mapping)
  - self-written mod.want_digest (instance digests following RFC 3230)
  
  ➔ Compatible with FTS3 and accessible via OAuth2 secured webDAV
  ➔ Transfers are possible between GRID storage and Apache
  ➔ Direct transfers between two Apaches are WIP
Transfer tests

- Endpoints:
  - DCache @ DESY as active party
  - Apache @ HZDR Dresden (OAuth2-secured webDAV EP) & @ DESY (+ local user mapping and instance digests) as passive parties

- Throughput of **20 to 100 MB/s**, depending on available bandwidth
- 1 TB of data could be transferred in roughly **3 - 15 h**
Summary

- Enabling data transfers between HGF centres with existing open-source software
- New software configuration for HTTP-TPC that is easy to deploy
- Together with FTS3 and WebFTS provides base for a transfer service
- Transfer tests between Apache endpoints and dCache instance successful
- Valuable addition to transfer services after module optimizations
- Currently limited to transfers involving grid storage endpoint (e.g., dCache), but future work will make Apache httpd itself capable of HTTP-TPC
- Practical solution for data transfers between Helmholtz centres
DANKE!
THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!

Grateful acknowledgements:
Andrea Manzi, EGI
Mihai Patrascoiu, CERN
Jan Erik Sundermann, KIT
backup slides
Endpoint components

- mod.want.digest (github.com/wetzel-desy/mod.want.digest): 
  - Developed by Tim Wetzel and Paul Millar, fragments taken from httpd's mod_negotiation
  - Implements instance digests in accordance with RFC 3230 (HTTP headers "Want-Digest" and "Digest")
  - Supports ADLER32, MD5 and SHA digests
  - Alpha version until now
    - No digest caching mechanism or on-the-fly calculation
    - Has to read file from disk for digest calculation
  - Good first version but needs to be optimized (WIP)