DSIT Report

Overview

DSIT WP1 WP2

Federated AAI and Federated Storage

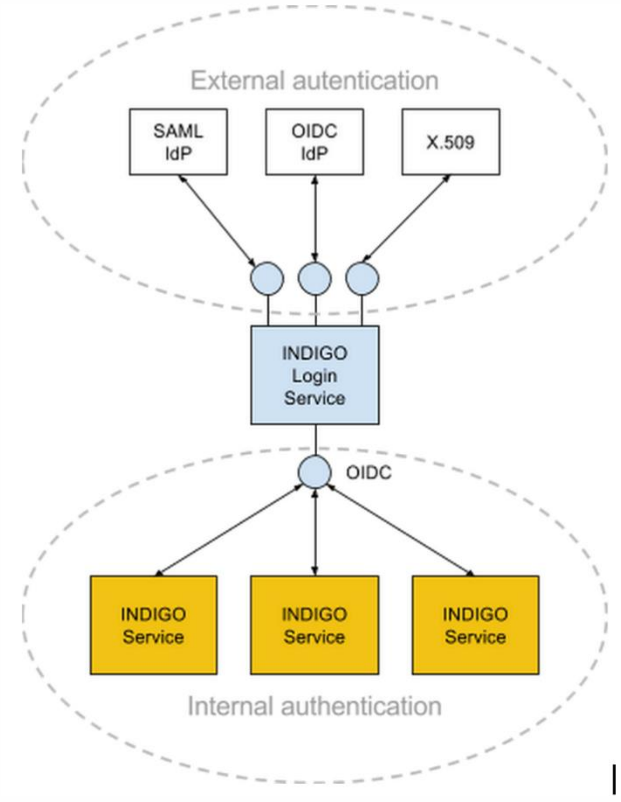
With contributions from

Marcus Paul Shiraz Patrick
Arsen GSI

WP1 Activities Overview

- Participation in INDIGO DataCloud (IDC)
- Colaboration with EUDAT
- Extensions of KIT LDAP Facade
- Pushing SAML and eduGAIN for research colaborations

Aligning LSDMA work with INDIGO-DataCloud



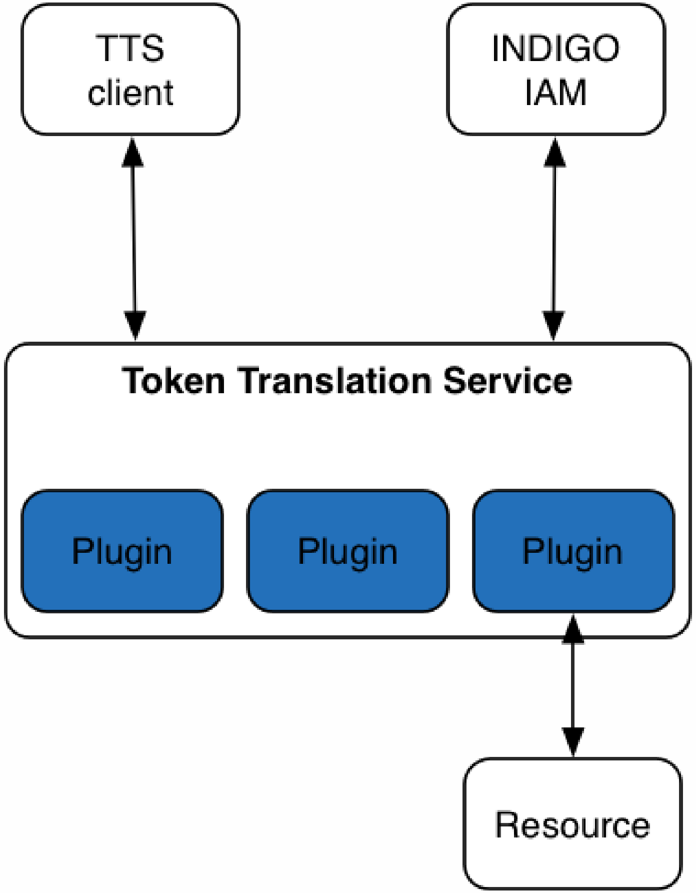
- Many ways that users can authenticate to **INDIGO DataCloud** (IDC) services:
  - SAML IdP → DFN AAI
  - X.509 will also be supported
- Clients authenticate with services using **OpenID-Connect** (OIDC).

  INDIGO Login Service is the „OIDC IdP“
- Clients get an **IDC token** that they use with INDIGO Services.

TTS: Support federated login to existing services

- For **Resources** that cannot use SAML or OpenID-Connect:

  e.g., GridFTP needs X.509, SSH server needs pub/priv key-pair
- One TTS service per site or service
- Client **authenticates** to TTS using INDIGO-DataCloud token
- TTS plugins:
  - Easy to (site-local) **customisation**
  - Can also do basic provisioning
  - supply **credential** to client for access to Resource
- Prototype ready by **end of March**.



Add support for INDIGO-DataCloud: dCache

- Initially targeting the **Relying Party** (RP) use-case.
  - **Useful for**: portal-, VM-, or WN- access to storage
  - **Not useful for**: direct user access.
- Initial support will be **generic OIDC**:
  - Good for INDIGO Tokens
  - Or: Google, Microsoft, Deutsche Telekom, …
  - Will be finished by **end of March**, will be part of dCache v2.16
- Future work includes:
  - INDIGO-specific **optimisation** (no network call-out)
  - Support for identity harmonisation (see next slide)
  - Direct user-access to storage using OIDC

KIT

- Identity Harmonisation
  - User has same identity whether logging in via SAML, OIDC or other (e.g., X.509)
  - Prototype working for
    - Google-account (OIDC)
    - Home-Institute-account (SAML)
  - Working on X.509
- KIT LDAP Facade (Accept many credentials to use various services)
  ([SSH|FTP|LOGIN] <=> [OIDC|SAML|X.509])
  - Prototype working for:
    - SSH with OIDC or SAML (plain password + SAML-token)
  - Working on
    - Robustness
    - Integration with IDH
    - X.509
  - Add support for guest users (here: via Umbrella from the synchrotron community)
    - Reduce set of required attributes from IdP
  - Demonstrator deployed at PSNC as an AARC pilot

KIT

- Macaroon Service (MacSer), protoype
  - Mint, add-caveats (1st + 3rd party), binding
- AARC
  - Requirements for policies and legal advice for international attribute release
  - Blueprint architectures for interoperable international FIM
    - Internal draft finalizing
    - Publication to AAI stakeholders at research infrastructures and projects
- On the roadmap:
  - Develop strategies for interoperable Levels of Assurance (LoA) (AARC, EUDAT)
  - Enable federated group management and secure data sharing (EUDAT)
    - Interface LDAP-Facade with external SAML attribute authorities
      - e.g.: Unity / b2access

GSI and AAI

- Design of new IdM system for GSI and FAIR
- LSDMA prototype for eduGAIN (wiki login) well received
- Solution currently under evaluation
  - Integration of users from GSI
  - How to ensure availability of GSI services to _all_ FAIR users
  - Once clarified: Recommendation for use in FAIR

FZJ

- UNITY update: current status
  - **pilot UNITY instance** to serve LSDMA and EUDAT:
    - Users can authenticate using their **home IdPs** (SAML / DFN-AAI),
    - Users can also authenticate using **„social“ IDs** (Google, Facebook, …)

      … they then **associate** this login with their federated ID.
  - LSDMA services can use UNITY as their **authentication service** (IdP):
    - works with both SAML (SP) or OAuth (RP)
    - The instance has recently been updated from v1.7.1 to v1.8.0

    (see http://www.unity-idm.eu/site/downloads for details)
- This pilot will remain **available and maintained** for the LSDMA (and EUDAT) users
- Integration with Idf under way

FZJ

- Future work:
  - **Integrate with UNITY**
    - Update SAML SPs and OAuth Clients within LSDMA
    - Which LSDMA SP/RPs do we want to target?
  - Target users with **different LoAs**:
    - Different services may have different requirements;
    - How to support this within the existing infrastructure.
  - Move towards **production readiness** of the service

=> Drop me an email so we can interact: hardt@kit.edu

WP2, Federated Storage

News :

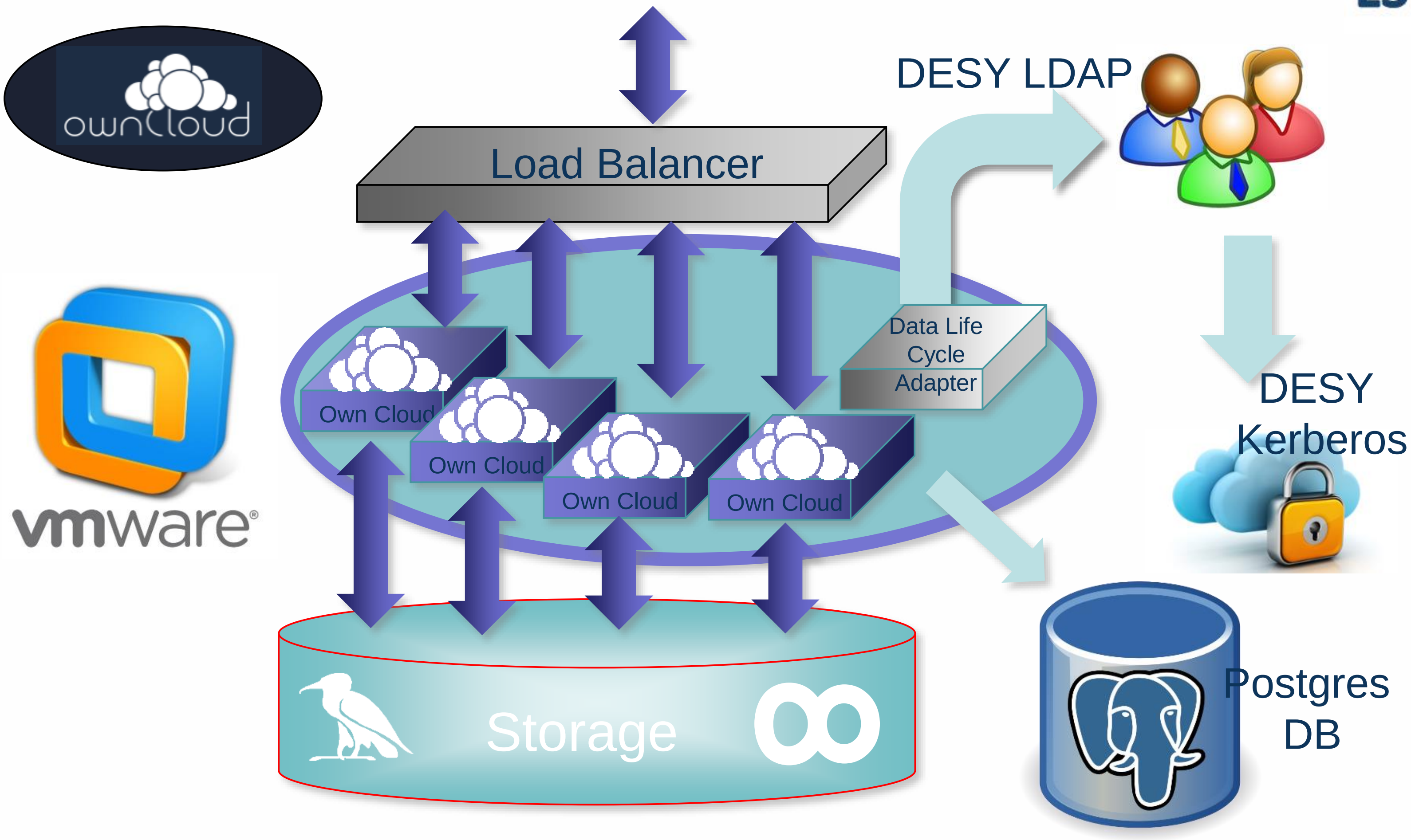
Attempt to improve the DSIT <=> DLCL interactions at DESY

- New LSDMA Position at DESY starting April or May 2016
  - dCache developers
  - Skills in XFEL data flow and storage policies
- Plan is to provide as much as possibile data management functionality for the XFEL data work flow by dCache.

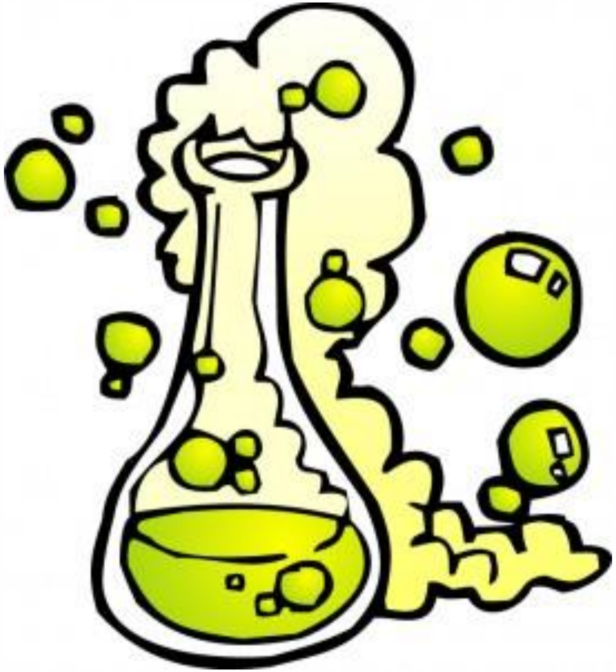
Services

- DESY Production service: OwnCloud – dCache hybrid
  - Provides sync’n share from OwnCloud and Multi Tier Support from dCache
- DESY Production service : DynaFed
  - Federating HTTP endpoints to a single overlay endpoint
- DESY Evaluation Service : WebFTS
  - GUI steered file transfer service (based on gridftp and http)
- ADD YOUR SERVICE HERE

Production system at DESY



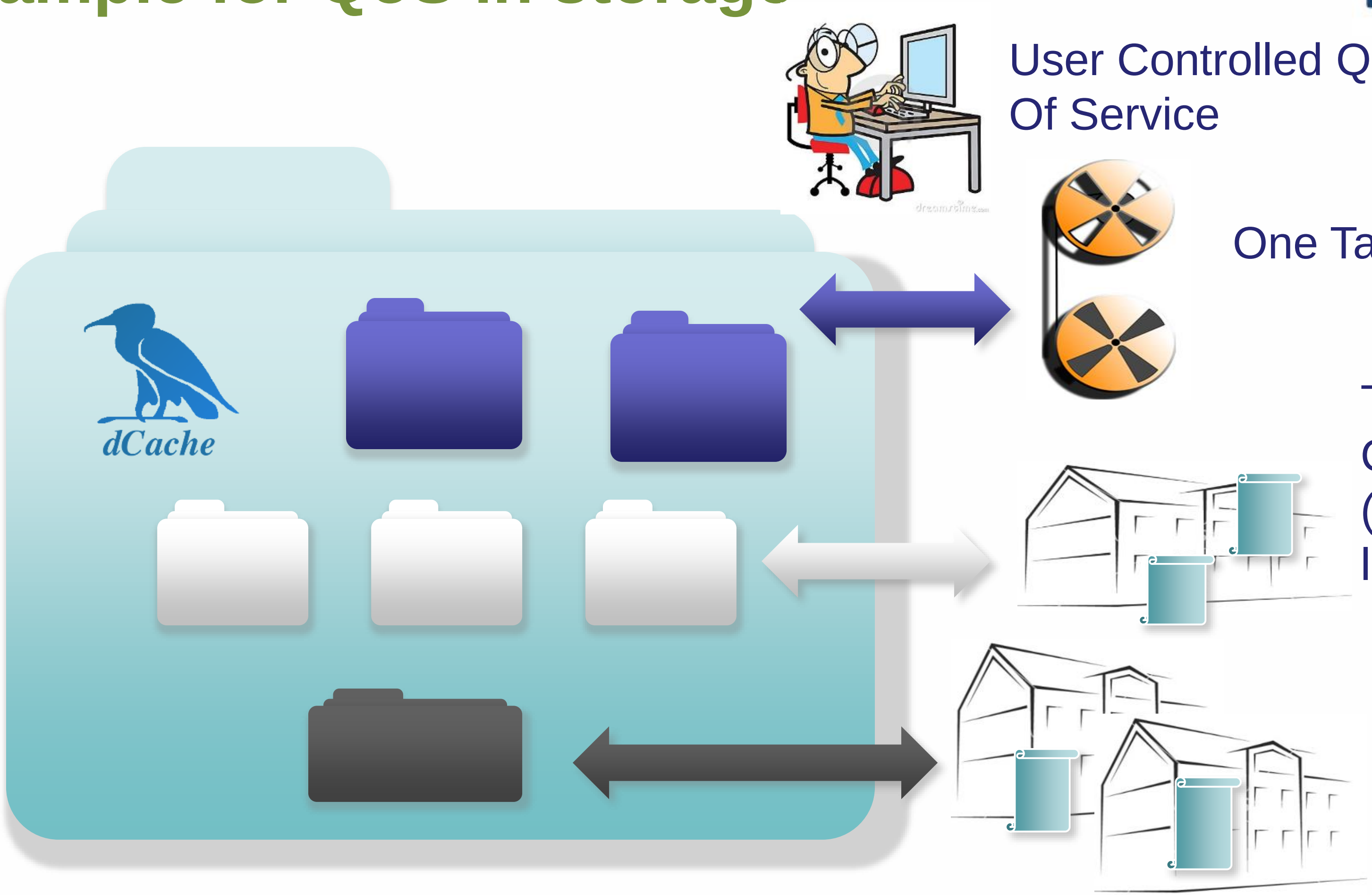
Current Projects



Ongoing Projects through INDIGO DataCloud

- Definition of QoS policies for storage within RDA (Working Group)
- Modelling the above Scheme within CDMI
- Providing a reference implemention for
  - GPFS / TSM by KIT
  - dCache by DESY
- Next Step: User defined Data Life Cycle

Example for QoS in storage



User Controlled Quality Of Service

One Tape Copy

Two Disk Copies (same location)

Two Disk Copies (different locations)

Ongoing dCache Projects

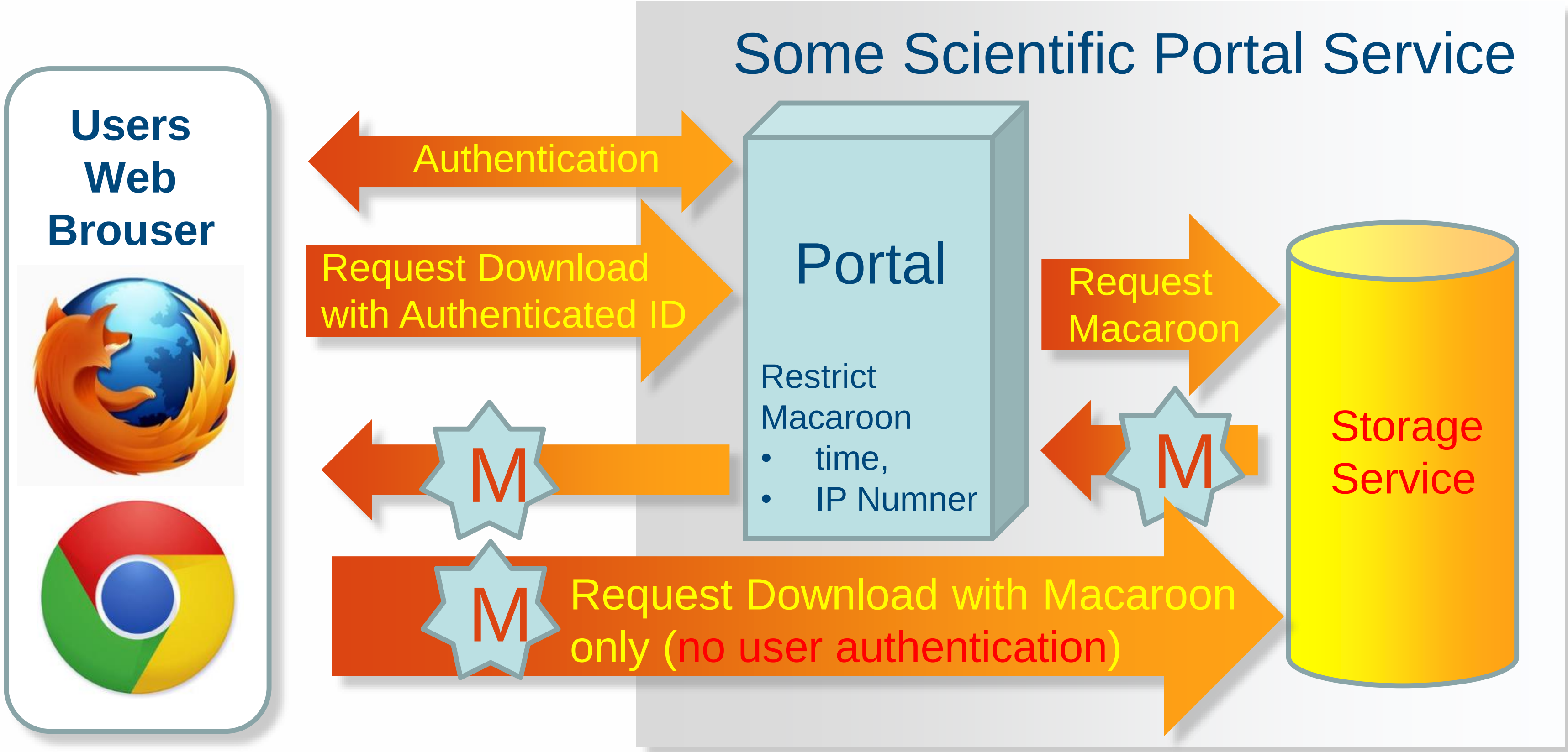
- Sync’n Share
  - Integration with OwnCloud (in production, see ‘service’ slides)
  - High Privicy Sync’n Share with DCORE ®
- Backend Filesystems
  - dCache In a Box (dCache Appliance) with DDN
  - dCache based on CEPH
- Software Defined Storage (QoS and DLC for INDIGO-DataCloud)
  - User controlled storage parameters
    - Latency
    - Retentions (high performance low latency -> archive)
    - Access control
  - User defined Data Life Cycle
    - Policy driven migration of storage parameters over time

Ongoing dCache projects (cont.)

- Authentication
  - Integrating “Open ID Connected”
  - Integrating Macaroons in dCache.(*)
- Federate dCache (Student HTW Berlin)
  - NDGF Model : One federated dCache
  - WLCG Model : Federation of dCaches resp. Dynamic Federations
- dCache in Docker containers (Student HTW Berlin)
- Improved dCache name space system based on no-SQL systems.
  - Collaboration with University Hamburg.
- Small files for Tape
  - dCache provides a pluggin for creating containers for small files to be scheduled for tape storage. (Improved tape system performance)

Macaroons (Example)

dCache handling Macaroons to support Portals



Some Scientific Portal Service

Users Web Brouser

Authentication

Request Download with Authenticated ID

Portal

Restrict Macaroon

- time,
- IP Numner

Request Macaroon

Storage Service

Request Download with Macaroon only (no user authentication)

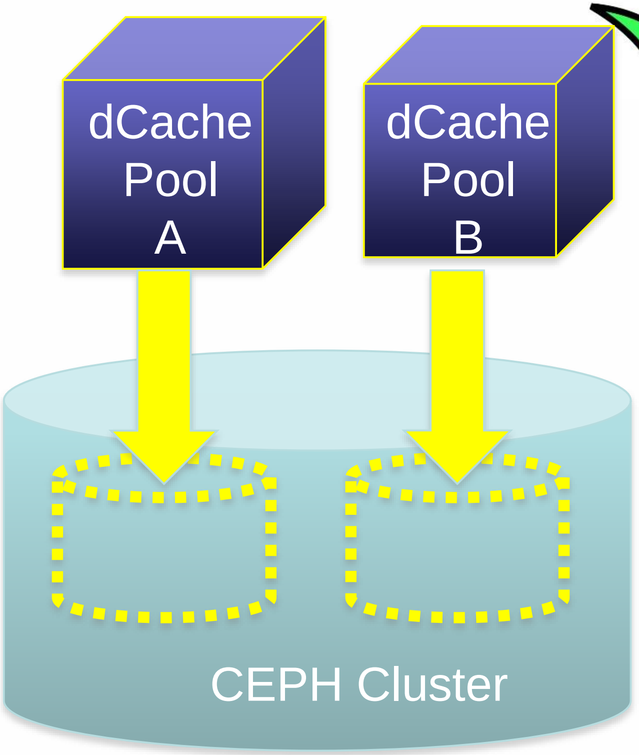
Details on ceph Integration

- CEPH complements dCache perfectly.
  - Simplifies operating dCache disks.
  - dCache accesses data as object-store anyways already.
- dCache is evaluating a ‘two step approach’.
  - Each pools sees it own object space in CEPH
  - All pools have access to the entire space, which is a slight change of dCache pool semantics.
- Would merge CEPH and dCache advantages
  - Multi Tier (Tape, Disk, SSD)
  - Multi protocol support for a common namespace.
    - All protocols see the same namespace
  - All the dCache AAI features
    - Support for X509, Kerberos, username/password

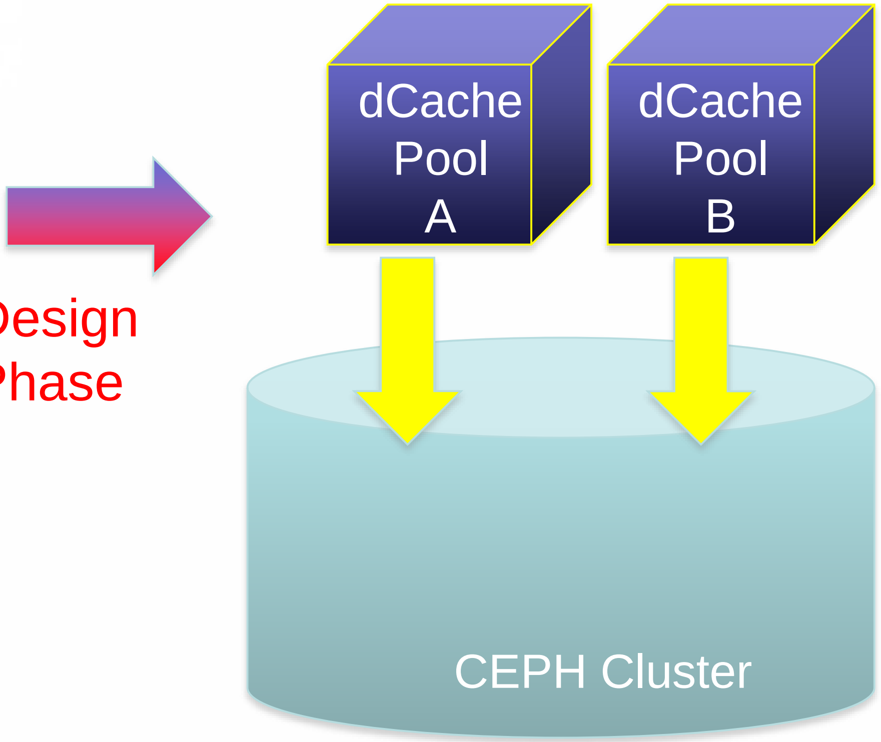
CEPH Integration Details

Each dCache pool still only ‘sees’ his own private repository.

DONE ✓



dCache pools can use shared repositories. Requires new pool semantics. (Focus on protocol engine)



Design Phase

Upcoming dCache Workshop

11 – 12 April, 2016
Barcelona,
UAB - CASA CONVALESCÈNCIA,
Visit www.dCache.org

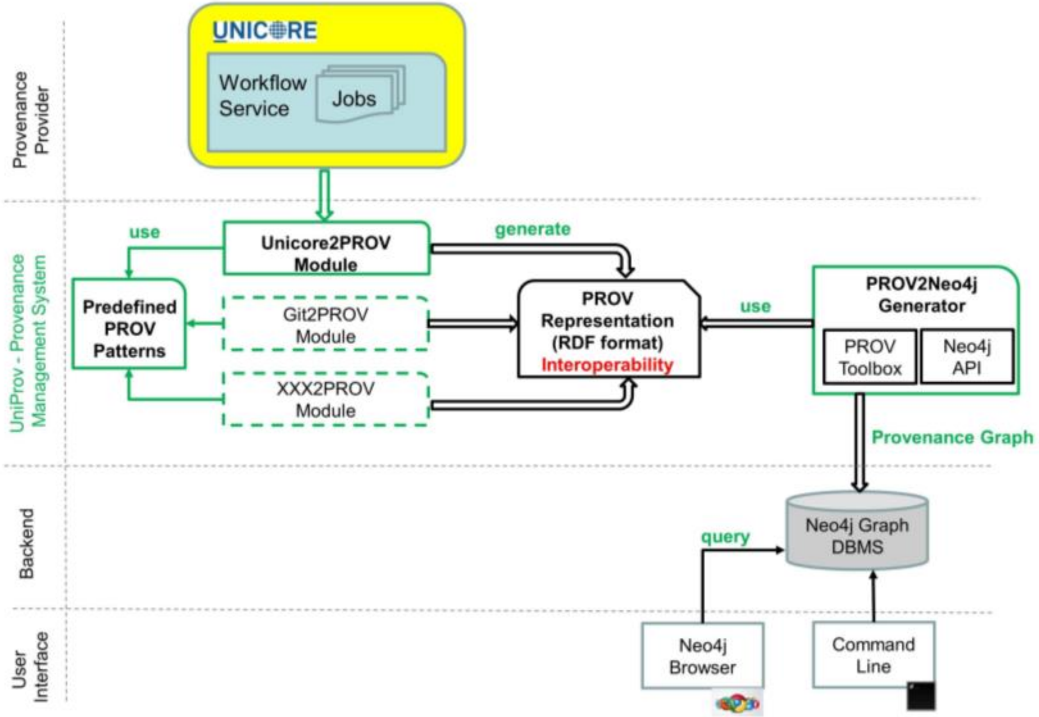
Data Services Integration Team

WP3 – Metadata Catalogues

Richard Grunzke, Bernd Schuller, André Giesler, Patrick Fuhrmann, Paul Millar, Stephan Kindermann, Volker Hartmann, Thomas Jejkal, Rainer Stotzka, Uğur Çayoğlu, Jörg Meyer, Ajinkya Prabhune

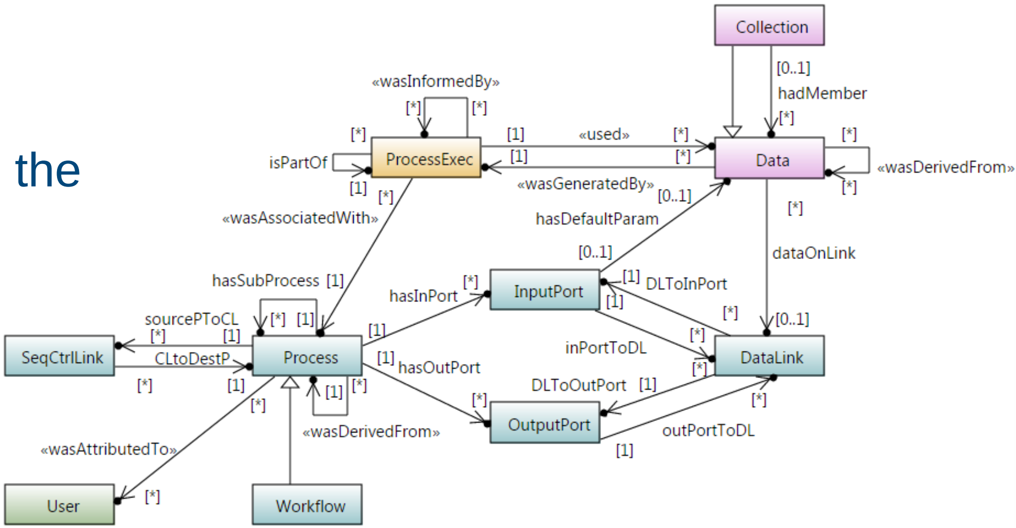
UNICORE – UniProv

- “A flexible Provenance Tracking System for UNICORE”
- UNICORE as provenance source & Neo4j graph db for visualization
- MASi integration planned



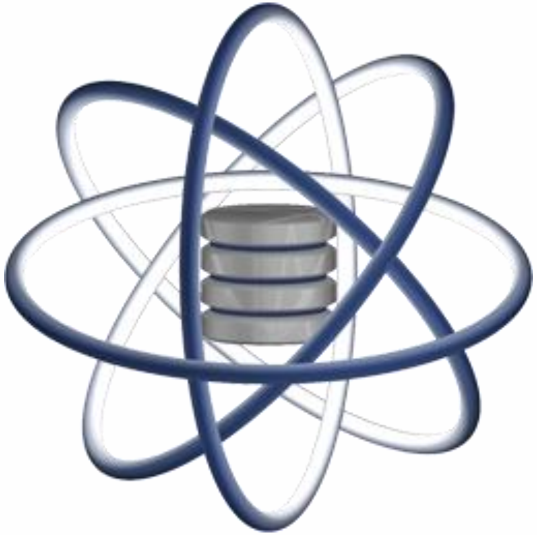
KIT – Provenance for Nanoscopy

- ProvONE provenance model
  - Graph based modeling of provenance (ProvONE)
  - Prospective & retrospective provenance
  - Interoperable with existing standards OPM/PROV
  - Graph database for persisting the ProvONE graphs
- Planned inclusion into KIT DM



KIT Data Manager - Development

- See WP6



http://www.kitdatamanager.net

Earth Environment at DKRZ

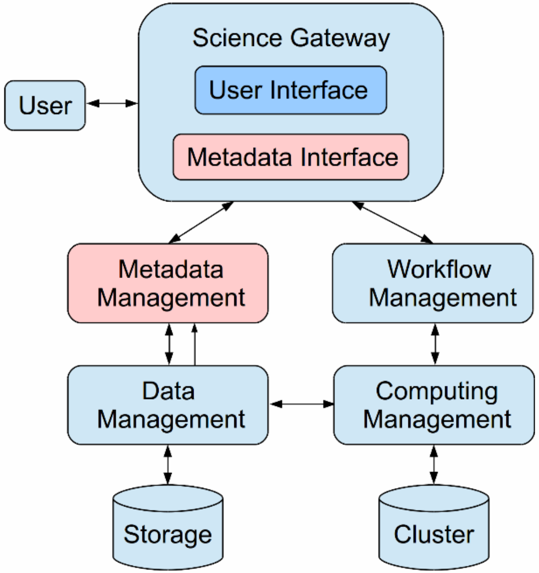
- Work on metadata-enabled Birdhouse web processing environment
- Prototype for (massive) PID and PID metadata generation to support next international climate model intercomparison project (CMIP6)
- Responsible for several recommendation papers for the world climate infrastructure panel (WIP)
- Work on EUDAT B2Find metadata search service continued

KIT - DKRZ PhD Metadata Cooperation

- PhD thesis on data management in climate research
- Integrated MongoDB instance for existing climate metadata (→ netCDF)
- Portal server for search (Flask & Plugins)
- Search applications via commandline or web interface

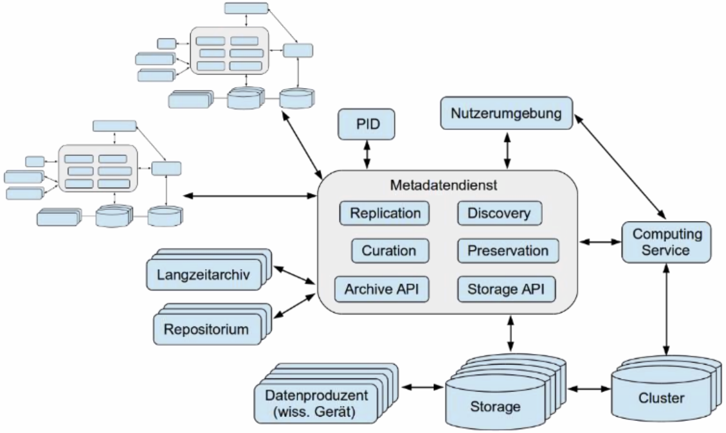
ZIH Metadata Research

- Coordination of work package
- Dissertation – Generic Metadata Handling Approach
  - Evaluation paper accepted in Journal of Grid Computing
  - Submitted thesis in December
  - Defense in April
- Coordination of MASi DFG project



MASi - Generic Research (Meta)data Service

- Requirement analysis finished
- Fixing overarching architecture
- Generic graphical interface work in progress
  - Liferay integration for common web capabilities
  - Generic web portlet with increasingly advanced integration and graphical capabilities
  - For quick community adaptations
  - Via low barrier of entry
- Generic interface for additional systems
- Towards first community prototype
- Work on additional external use cases

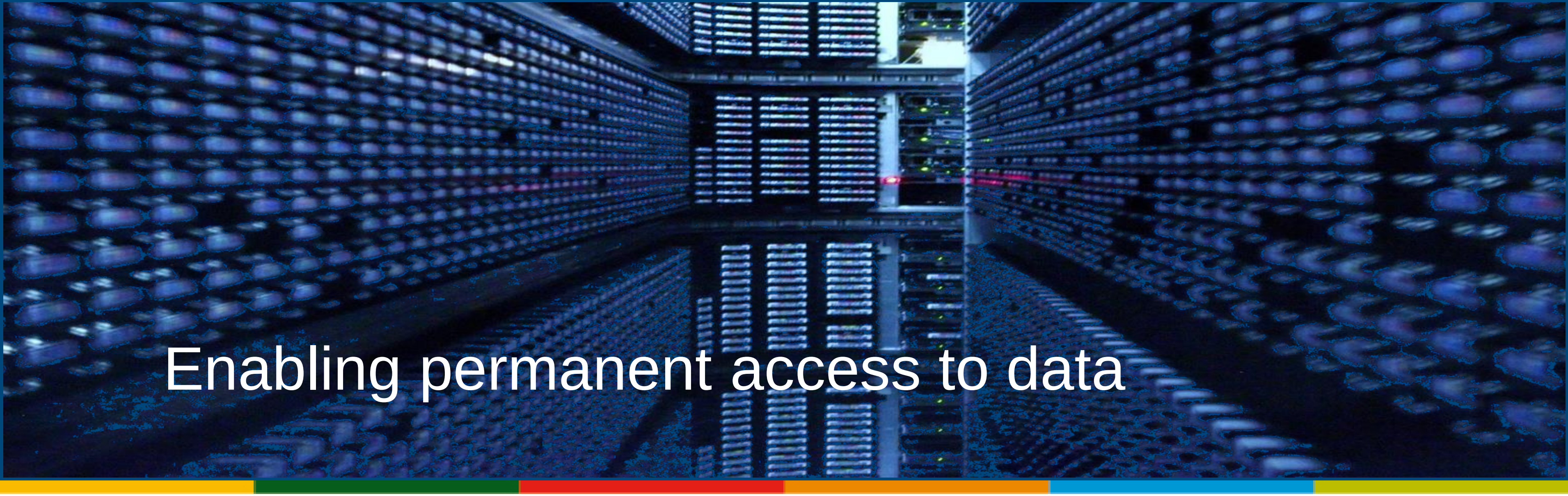


Publications

- R. Jäkel, R. Müller-Pfefferkorn, M. Kluge, R. Grunzke and W. E. Nagel: Architectural Implications for Exascale based on Big Data Workflow Requirements, Big Data and High Performance Computing, IOS Press, 2015, 26, 101 - 113.
- J. Krüger, P. Thiel, I. Merelli, R. Grunzke and S. Gesing: Portals and Web-based Resources for Virtual Screening, Current Drug Targets, http://www.eurekaselect.com/node/138922/article, 2016.
- R. Grunzke, J. Krüger, R. Jäkel, W. E. Nagel, S. Herres-Pawlis and A. Hoffmann: Metadata Management in the MoSGrid Science Gateway - Evaluation and the Expansion of Quantum Chemistry Support, Journal of Grid Computing, accepted.

DSIT WP4 - Archives

All hands meeting, March 2016



Enabling permanent access to data

Long time storage : business cases from 2013

A. Project store
- Temporary storage for large data, 3 to 4 years
- Non active data that active experiments depend on
- Simple upload-interface

~ EUDAT B2SHARE

B. Good scientific practice store, long term storage
- All of **A** above plus
- At least 10 years, conformity with good scientific practice
- Minimal set of meta data, PID for reference
- Expiration, Integrity checks

~ EUDAT B2SAFE

bwDataArchiv

C. Archive aka repository
- All of **B** above plus
- Forever or as long as is needed
- Interaction with community
- Data curation management
- Premium: Conformity with OAIS, certification of repository (DSA, ISO16363)

OpARA

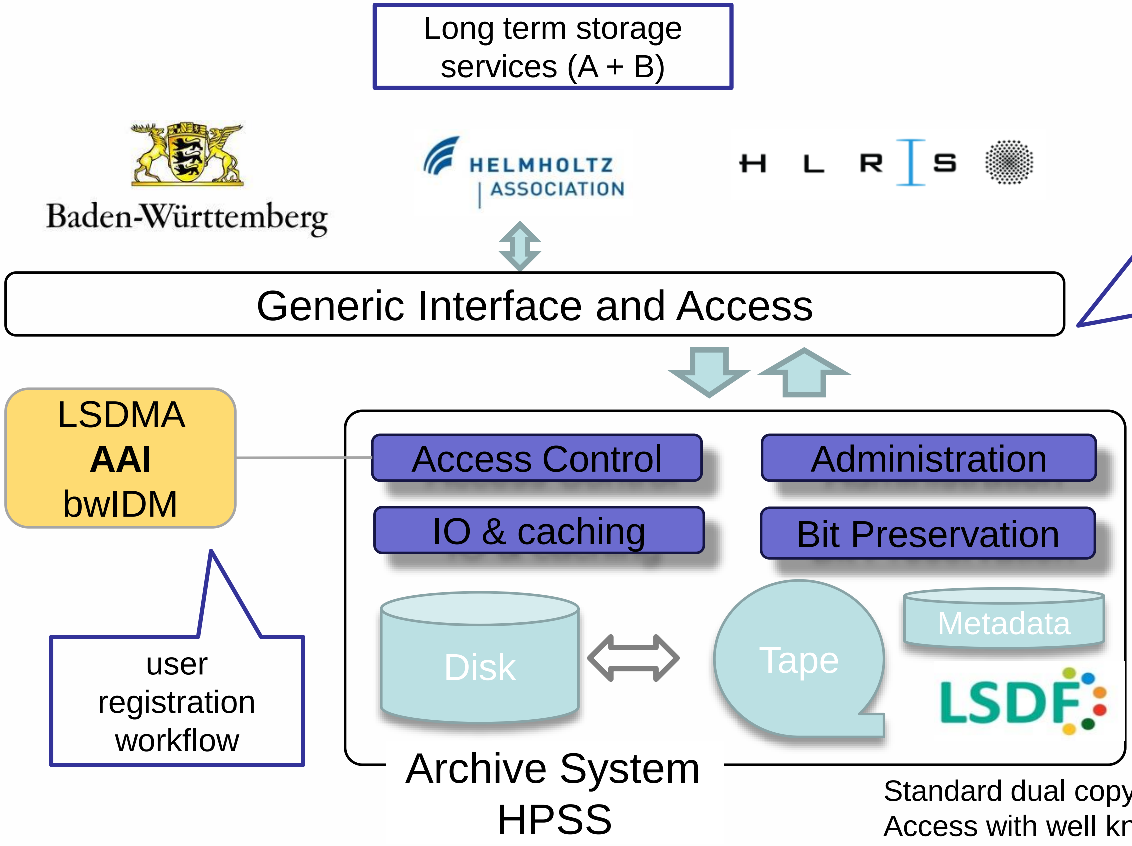
RADAR

TU – Dresden

- OpARA - Open Access Repository and Archive
  - In implementation phase (based on DSpace)
  - institutional repository with long term preservation
  - Shibboleth authentication
  - agreement on basic metadata for data sets
  - currently progresses on first prototype → user testing

Note: German DSpace User Group Meeting to be Held in Hamburg, Sept. 27, 2016

Archives for scientific data



collaboration of LSDMA WP4 with EUDAT, RADAR, bwDIM definition of API for Archives:

+ checksumming
+ ‘technical’ metadata
+ asynchronous access
+ storage quality / costs

INDIGO - DataCloud

bwDIM

EUDAT

RADAR

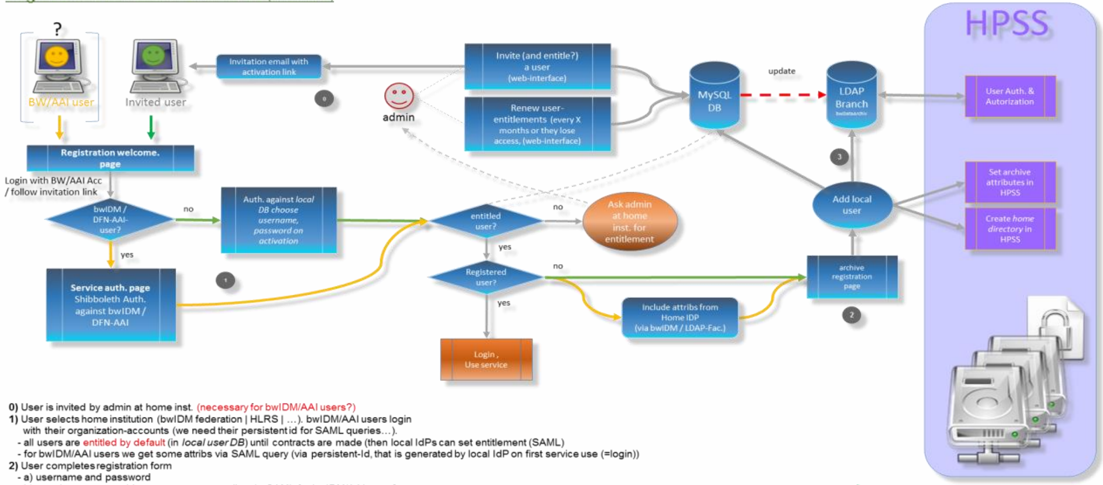
Standard dual copy, geolocation separated storage
Access with well known clients (SFTP and GridFTP/GT)

BWDATA ARCHIV

User registration workflow (development in bwDIM and bwDA)

- long term registration
  - personal account / no transfer of ownership
- supervisor role for groups and datasets
  - authentication hooks for community data curator
- ‘linking’ of previously archived data with new account
- access for entitled users (BW)
  - integration with bwIDM
- other users / third parties
  - workflow to authorize new entitlement
  - access after title reception
- collaboration with WP1

Registration for Archive Service V3 (25.02.2016)


?
BW/AAI user
Invited user
Invitation email with activation link
0
admin
Invite (and entitle?) a user (web-interface)
Renew user-entitlements (every X months or they lose access, (web-interface)
MySQL DB
update
LDAP Branch
HPSS
User Auth. & Autorization
Registration welcome. page
Login with BW/AAI Acc / follow invitation link
bwIDM / DFN-AAI-user?
no
yes
Auth. against local DB choose username, password on activation
Service auth. page Shibboleth Auth. against bwIDM / DFN-AAI
1
entitled user?
no
yes
Ask admin at home inst. for entitlement
Registered user?
no
yes
Include attribs from Home IDP (via bwIDM / LDAP-Fac.)
Login , Use service
3
Add local user
archive registration page
2
Set archive attributes in HPSS
Create home directory in HPSS

0) User is invited by admin at home inst. (necessary for bwIDM/AAI users?)
1) User selects home institution (bwIDM federation | HLRS | …). bwIDM/AAI users login
with their organization-accounts (we need their persistent id for SAML queries…).
- all users are entitled by default (in *local user DB*) until contracts are made (then local IdPs can set entitlement (SAML)
- for bwIDM/AAI users we get some attribs via SAML query (via persistent-Id, that is generated by local IdP on first service use (=login))
2) User completes registration form
- a) username and password
- b) org., inst., repr., orcid etc, we get some attribs via SAML for bwIDM/AAI users]
3) User is added to ‘local’ DB (MySQL / own LDAP)

DEPROVISIONING:
- *Extern users* that are not “renewed” by *admins* loose entitlement in local DB autom. after x months (account and data is kept) (both loose write access only ?)
- *bwIDM/AAI user accounts* are “deprovisioned” as soon as a (periodical) SAML-query returns “account is deactivated” (attribs not returned->which?)…

➡ Unregistered ext. users (e.g. HLRS)
➡ Unregistered BW/AAI users
➡ All users

Service integration

- Archive storage interface design (with RADAR and EUDAT)
  - Hashes/checksums (EUDAT)
    - prototype in development
    - integration with iRODS and HPSS planned
  - Support for technical / administrative (PREMIS based) metadata
    - prototype with REST-full interface and DB is ready
    - integration with HPSS planned
  - Storage classes
    - single, dual, triple (shelf)
- Long time storage service (BW Projects)
  - Authentication
    - development of user registration workflow
    - integration with Archive services planned (Summer 2016)

Data Services Integration Team

WP5 – Performance and Power Optimization

Checkpoint Compression: Motivation

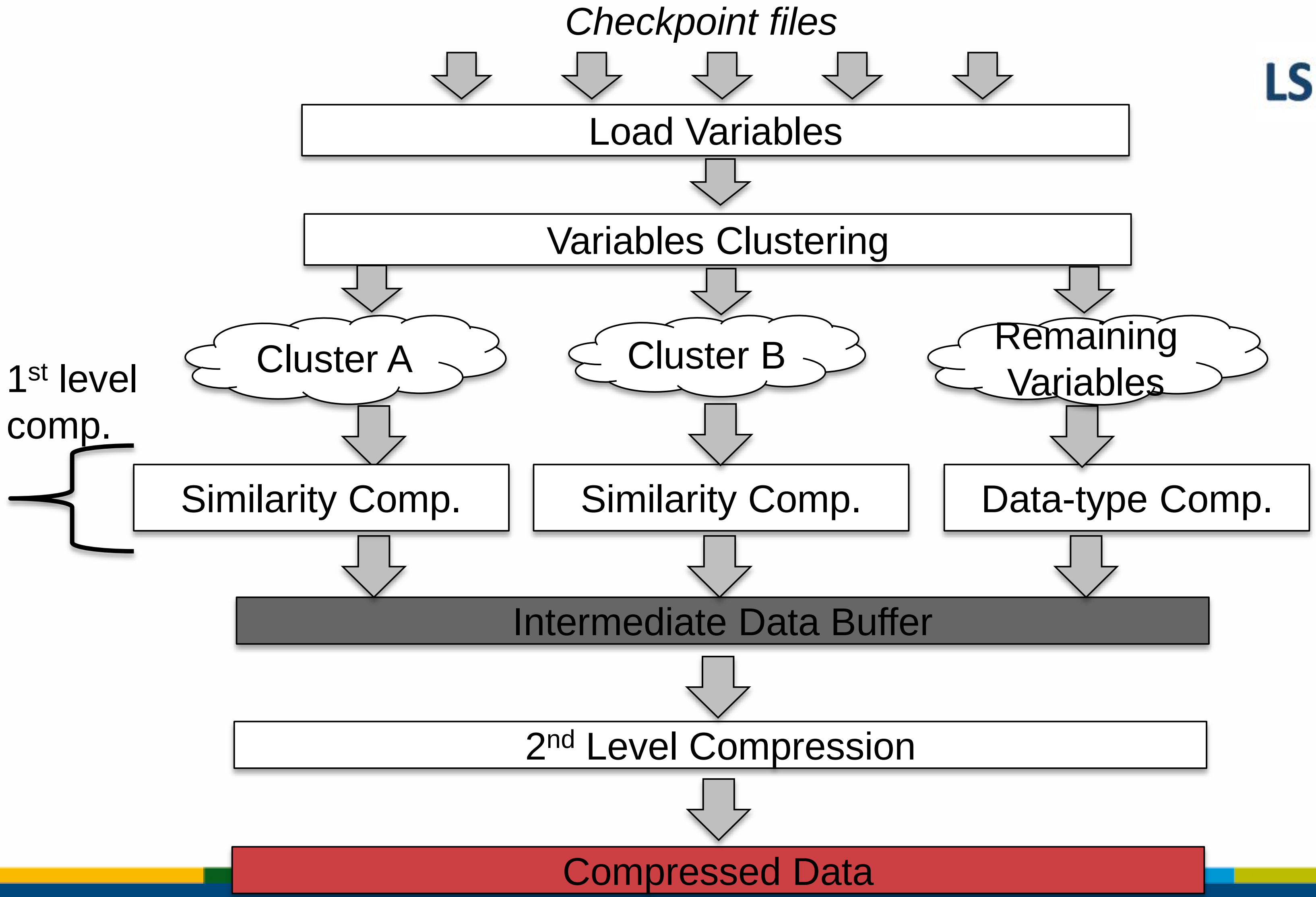
- Scientific applications deal with vast amounts of data
    - Storage needs continuously increasing
    - Require high speed/bandwidth link between computer and storage sub-system
- Use rich file formats that allow data annotation
    - Data type, variable name etc.
    - Variables can be scalar or arrays
- Elements of variables may have numerically close values

Checkpoint Compression: Goals

- Reduce footprint of scientific data by leveraging similarity in variables within a single file and across multiple files
- Research questions to address
  - Which algorithm to use for clustering variables?
    - Define a suitable distance/similarity metric
  - Which algorithm for compressing similar data?
    - Leverage properties implied by the similarity metric

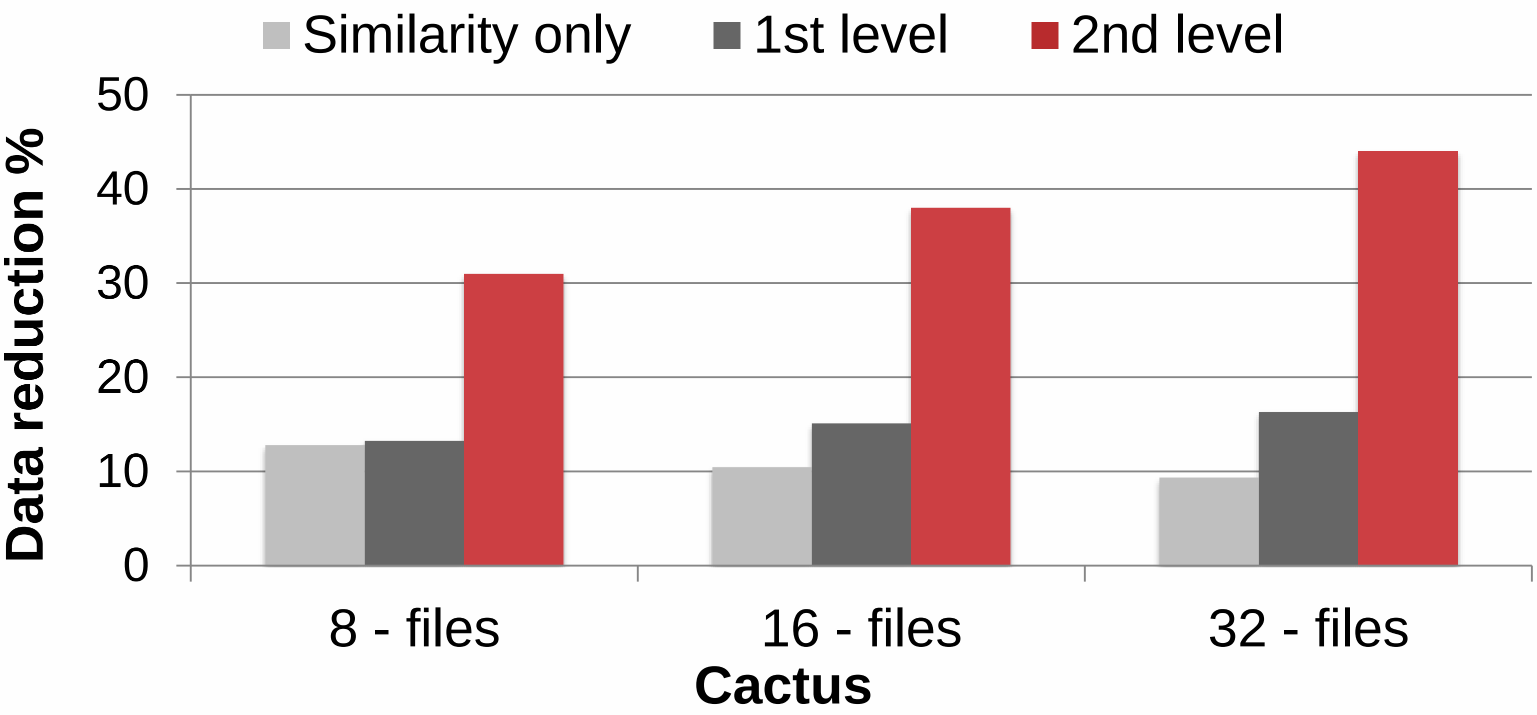
Checkpoint Compression: Scheme

- First level compression:
  - Similar variables with “similarity aware compression”
  - Concatenate rest of the variables and apply data-type-aware compression algorithms: fpzip, fpc etc.
- Second level compression:
  - Compress the output buffer of first level compression using generic compression algorithms: gzip, bzip2 etc.


Checkpoint files
Load Variables
Variables Clustering
Cluster A
Cluster B
Remaining Variables
1st level comp.
Similarity Comp.
Similarity Comp.
Data-type Comp.
Intermediate Data Buffer
2nd Level Compression
Compressed Data

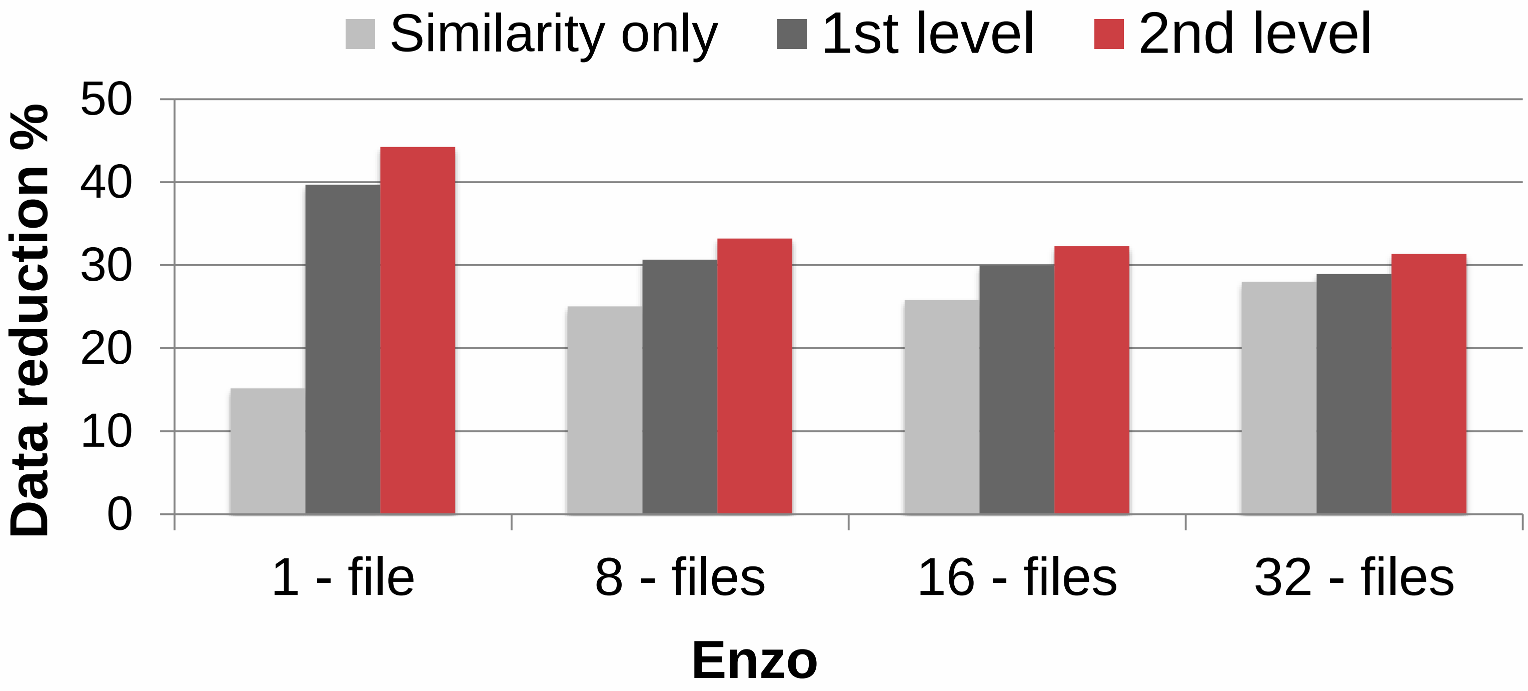
Checkpoint Compression: Preliminary Results

Volume reduction in each level of compression



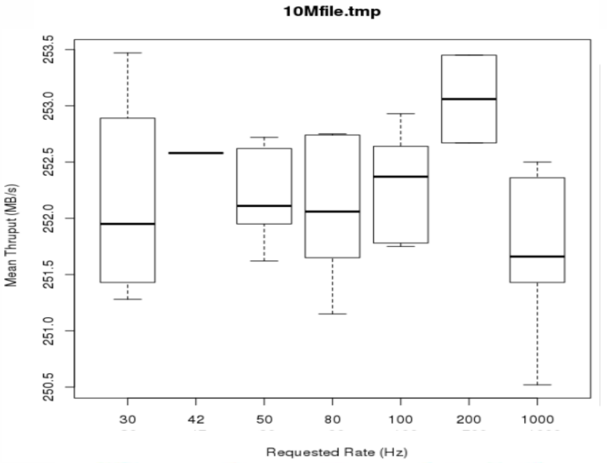
Checkpoint Compression: Preliminary Results

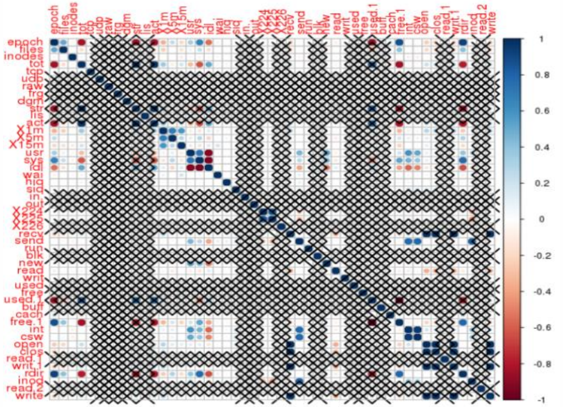
Volume reduction in each level of compression



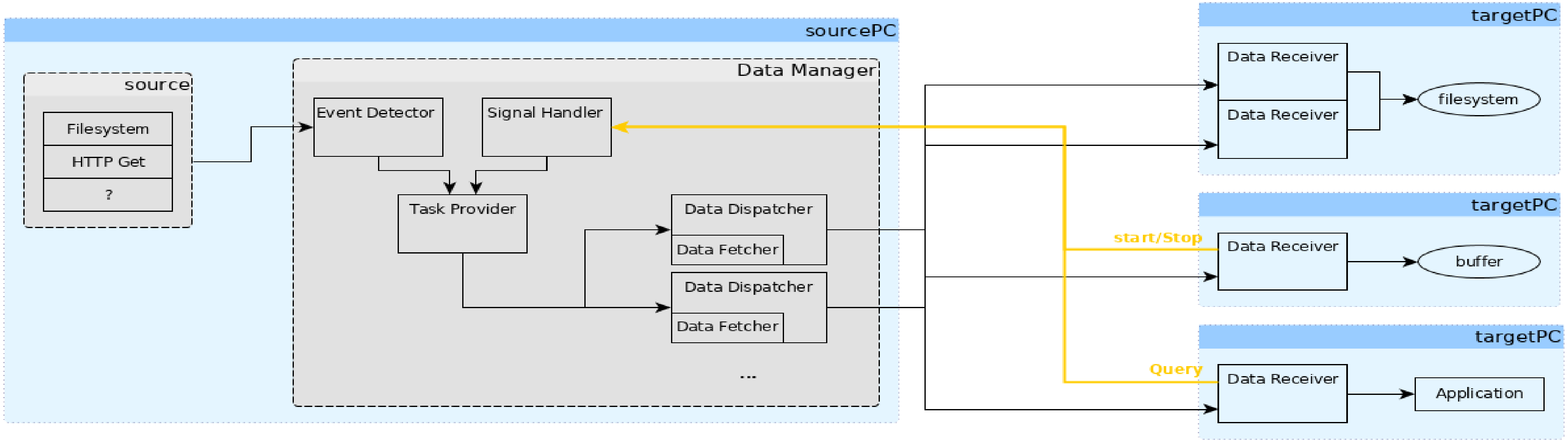
Automatic Performance Validation Tool

- Validates NFS and SMB connections from detector nodes to a storage system
- Fake data are created by a dedicated simulation tool, generates data of given size and rate
- Tests are done on a regular basis in order to find potential problems as soon as possible
- Tests compare the reference performance data set with the current status
- Different type quantities are collected to measure correlations and dependencies
- The tool is in development phase





ZeroMQ Data Transfer



Available event detectors:

- Based on inotifyx library (Linux)
- Based on watchdog library (Linux/Windows)
- Get events via an API through ZeroMQ

Available data fetcher:

- Read from file system
- Get data via an API through ZeroMQ

→ easily expandable

Available receiver types:

- Store as files
- Forward to an application
- Build HDF5 (in development)

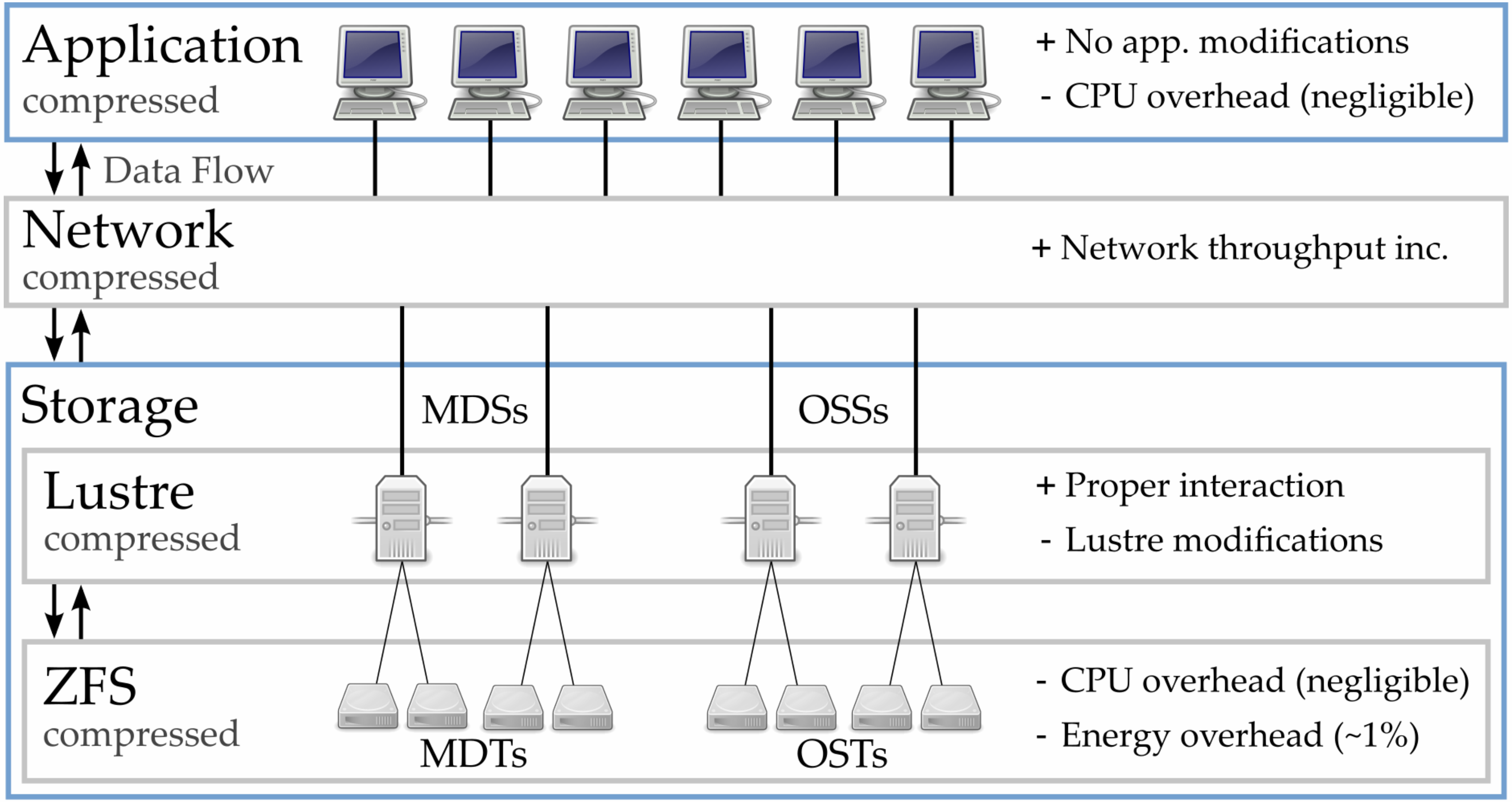
ZeroMQ Data Transfer

- Based on Python and ZeroMQ
- Data can be distributed to different targets with different requirements in parallel
  - E.g. to store the data in the storage system directly
- A generic tool set for different data streams with different QoSs
- Modular architecture:
  - Event detectors (for different libraries, for applications)
  - Data fetchers (from filesystem, directly from application)
  - Receivers

Enhanced Adaptive Compression in Lustre

- Part of our IPCC for Lustre
- Proper support for compression in Lustre
    - Interaction with application-specific compression
- Support at different levels
    - Servers, clients and within applications
    - Completely transparent to applications (tuning via ladvise)
- Compression should be adaptive
    - Based on information about the data, the current load etc.

Enhanced Adaptive Compression in Lustre



Compression Study

- Analyze scientific data from different domains
  - Different algorithms necessary for adaptive compression
  - Large scale study using various algorithms, file types and data sets
- Performance, compression ratio etc. for each algorithm
  - Behavior with incompressible data is also important
- Provide suggestions for data centers and researchers

- If we can get access to your data: get in contact!

Checkpoint Compression: Future Work

- Improve similarity metric/compression algorithm
  - Work on parts of the variable instead of all values
- Evaluate data from different applications
- Explore other similarity metrics
- Experiment with different clustering algorithms
- Parallelize current implementation

Publications

- OnDA: Online Data Analysis and Feedback for Serial X-Ray Imaging (Valerio Mariani, Andrew Morgan, Chun Hong Yoon, Thomas J. Lane, Thomas White, Christopher O'Grady, Manuela Kuhn, Steve Aplin, Jason Koglin and Henry N. Chapman), In Journal of Applied Crystallography
- MPI-Checker - Static Analysis for MPI (Alexander Droste, Michael Kuhn, Thomas Ludwig), In Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, LLVM '15, ACM (New York, USA), SC15, Austin, Texas, USA, ISBN: 978-1-4503-4005-2, 2015-11
- Big Data Research at DKRZ – Climate Model Data Production Workflow (Michael Lautenschlager, Panagiotis Adamidis, Michael Kuhn), In Big Data and High Performance Computing (Lucio Grandinetti, Gerhard Joubert, Marcel Kunze, Valerio Pascucci), Series: Advances in Parallel Computing, Edition: 26, pp. 133–155, IOS Press, ISBN: 978-1-61499-582-1, 2015

Data Services Integration Team

WP6 - Data Intensive Computing

Thomas Jejkal, Bernd Schuller, Daniel Becker, Pavel Efros, Richard Grunzke

Infrastructure and Service (1)

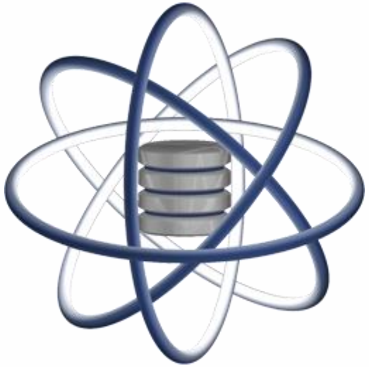
- Releases of UNICORE 7.4 and 7.5
  - YARN + HDFS support implemented
  - CDMI storage backend support

- TSI daemon rewritten in Python
  - More flexible storage configurations

- REST API for workflow submission and management

Infrastructure and Service (2)

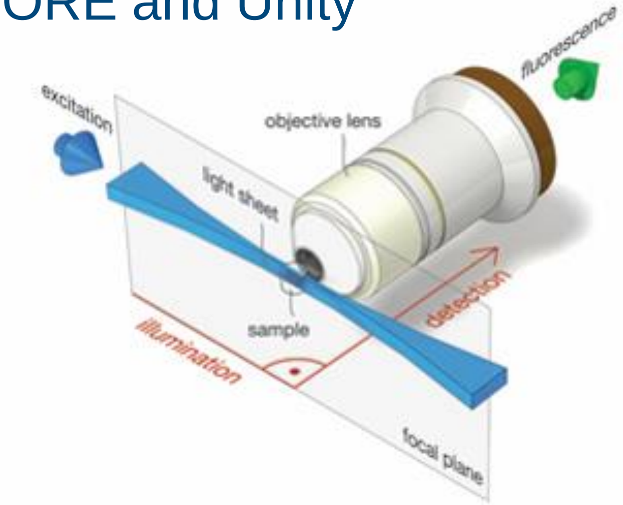
- Releases of versions 1.2
  - Enhanced metadata handling module added
  - Support for Digital Object transitions and types
  - Performance improvements for huge repository instances



- Public release of Data Workflow Service
  - Defines generic processing workflow
  - Allows to build up DAG processing chains
  - Current implementation supports local execution

High Throughput Microscopy

- Working on CDMI support in close cooperation with UNICORE and dCache
  - CDMI supported by UNICORE since version 7.5
  - Support for dCache ongoing work
- Pre-Production certificate-free HPC via UNICORE and Unity



Algorithm Development (HTW Berlin)

- Thesis “Realtime Analysis of Large-Scale Data” has been written and submitted
- Algorithms for categorizing images and localizing signals are discussed
- A prototypical implementation is reviewed in terms of runtime and efficiency
- A feasible solution for handling the vast amount of data created by a new generation of photon science experiments is proposed
    - Reducing the flood of data by accelerator devices (FPGAs/GPUs)
        - Located near the detector for vetoing “empty” diffraction images in realtime
        - More thorough analysis in near-realtime (as a second step)

Publications

R. Jäkel, R. Müller-Pfefferkorn, M. Kluge, R. Grunzke and W. E. Nagel: Architectural Implications for Exascale based on Big Data Workflow Requirements, Big Data and High Performance Computing, IOS Press, 2015, 26, 101 - 113.

J. Krüger, P. Thiel, I. Merelli, R. Grunzke and S. Gesing: Portals and Web-based Resources for Virtual Screening, Current Drug Targets, http://www.eurekaselect.com/node/138922/article, 2016.

R. Grunzke, J. Krüger, R. Jäkel, W. E. Nagel, S. Herres-Pawlis and A. Hoffmann: Metadata Management in the MoSGrid Science Gateway - Evaluation and the Expansion of Quantum Chemistry Support, Journal of Grid Computing, accepted.

P. Efros, E. Buchmann, A. Englhardt and K. Böhm: How to Quantify the Impact of Lossy Transformations on Event Detection, IEEE Transactions on Knowledge and Data Engineering (TKDE)

G. Hollmig, M. Horne, S. Leimkühler, F. Schöll, C. Strunk, P. Efros, E. Buchmann and K. Böhm: An Evaluation of Combinations of Lossy Compression and Change-Detection Approaches, Informations Systems

D. Becker: Dissertation mit dem Thema Reatime-Analyse großskaliger Datenmengen, 2016

P. Efros: Dissertation mit dem Thema Lossy Time-Series Transformation Techniques in the Context of the Smart Grid, 2016

Future Work in WP6

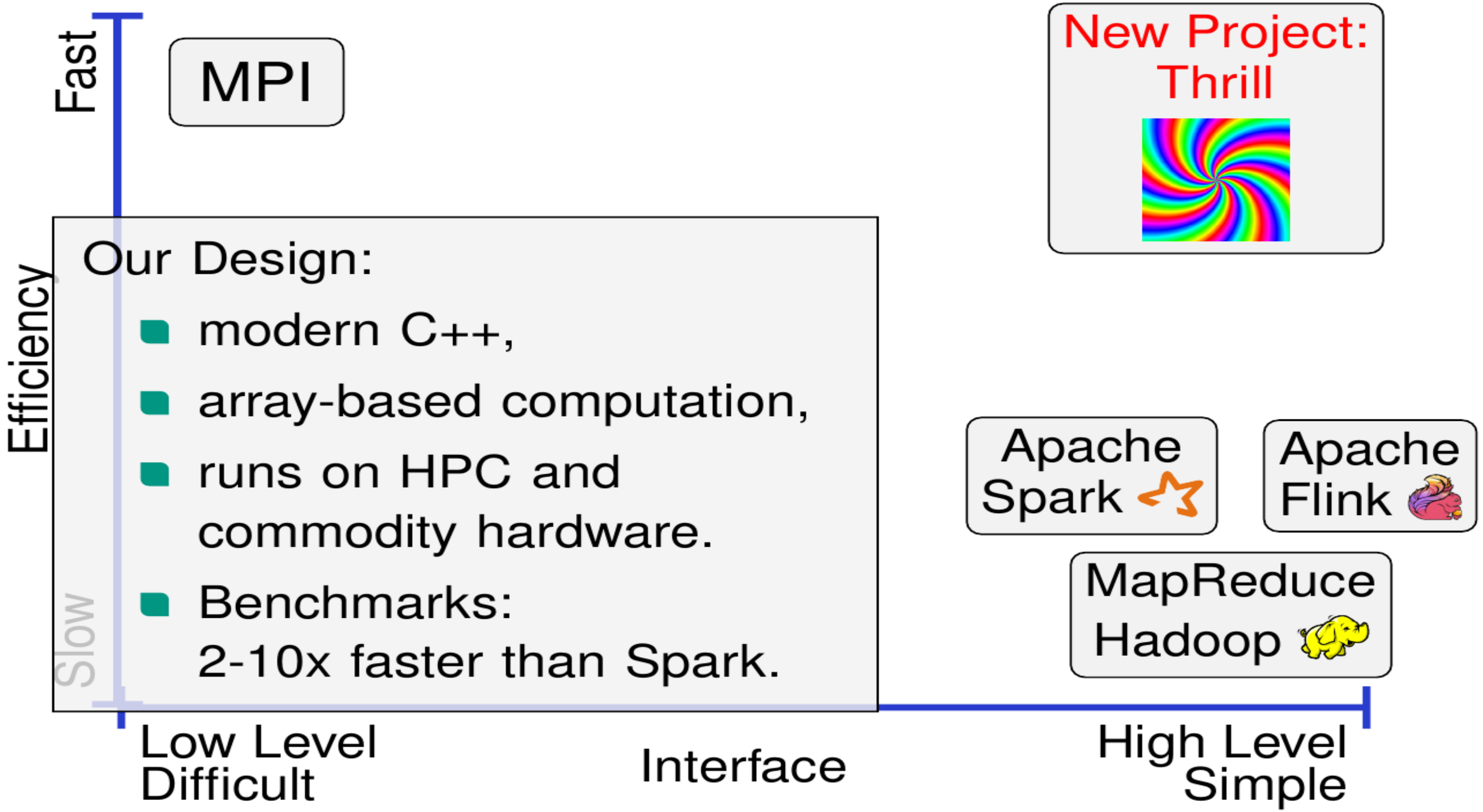
Infrastructure and Basic Services (1)

- Planned features for 7.6, 7.7
  - Improve existing OpenStack support
  - Failover/clustering of UNICORE servers for large-scale deployments
  - Python Client API
  - Secure interaction with Docker
  - Job execution on Amazon EC2 systems

Infrastructure and Basic Services (2)

- Implementation of Data Workflow Service use cases
    - NORDR, CodiHub, MASi
    - Evaluation and bugfixing
- Evaluation (and realization?) of a UNICORE integration into Data Workflow Service
    - Relevant for DFG project MASi
    - Basic single job submission

Thrill - Big Data Processing in C++



Our Design:

- modern C++,
- array-based computation,
- runs on HPC and commodity hardware.
- Benchmarks: 2-10x faster than Spark.