Overview: FIDIUM at Goethe University

14/02/2023

A. Redelbach

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Themenbereich 1: Entwicklung von Werkzeugen zur Einbindung heterogener Ressourcen - FIDIUM overview

Arbeitspaket 1: Erschließung und effiziente Einbindung von opportunistischen Ressourcen

Arbeitspaket 2: Accounting und Controlling von heterogenen Ressourcen

FIDIUM

October 2021 – September 2024

Standort	PI	FTE	Experiment	AP 1	AP 2
KIT	G. Quast / A. Streit	0.66	CMS	X	X
U Bonn	P. Bechtle	1	ATLAS/Belle II	X	
GU Frankfurt	V. Lindenstruth	0	ALICE/CBM	X	
U Freiburg	M. Schumacher	1.2	ATLAS	X	X
U Göttingen	A. Quadt	0.5	ATLAS	X	
U Wuppertal	C. Zeitnitz	0.5	ATLAS		X
Assoziiert					
GSI	K. Schwarz	-	ALICE	X	
DESY	V. Gülzow	-	verschiedene	X	
GridKa	A. Petzold	-	verschiedene	X	X

Themenbereich 1: Milestones at Frankfurt

AP I-1

- M1: Performance-Vergleichsmessungen von Laufzeiten repräsentativer Rekonstruktionsalgorithmen bei Prozessierung auf CPUs und GPUs (Q3/2022)

- M2: Optimierung der Prozessierungsschritte repräsentativer Rekonstruktionsalgorithmen bei Aufteilung auf CPUs und GPUs zur Reduzierung der Gesamtlaufzeit bei effizienter Ressourcennutzung (Q3/2023)

- M3: Container-basierte Lösungen zur effizienten Nutzung der entwickelten, optimierten CPU/GPU-Prozessierung in einem weiteren Experiment oder einem weiteren Standort (Q2/2024)

Themenbereich 1: Frankfurt status/plans

Detailed study of performances of reconstruction codes on CPUs/GPUs:

- Different architectures
- Tasks for local reconstruction
- Track fitting with optimal loading of threads
- Scalability for growing size of input datasets

→performance of each device can vary significantly depending on settings (e.g. CPU affinity) and the size of the dataset

Set of benchmarks for CPU/GPU performances in reconstruction tasks

Integration in other computing environments to be discussed

Goal: Efficient utilisation of CPU and GPU resources in compute intensive workflows

More details: → Presentation at FIDIUM annual meeting 2022
→ Talk by Grigory Kozlov today

Themenbereich 2: Data-Lakes, Distributed Data, Caching - FIDIUM overview

Arbeitspaket 1: Aufbau eines Echtzeit Data-Lake-Monitoring-Systems

Arbeitspaket 2: Technologien für Data-Lake-Caching

Arbeitspaket 3: Technologien für Data-Lake-Daten- und Workflow-Management

Arbeitspaket 4: Data-Lake-Prototypen, Technologien für QoS und effiziente Anbindung

Standort	PI	FTE	Experiment	AP 1	AP 2	AP 3	AP 4
KIT	G. Quast / A. Streit	0.66	CMS		X	X	
KIT	R. Engel	0.5	Auger/Einst.-Tel./IceCube	X		X	X
GU Frankfurt	V. Lindenstruth	1	ALICE/CBM		X		X
U Mainz	F. Maas / A. Brinkmann	1	PANDA		X	X	X
LMU München	G. Duckeck	1	ATLAS		X		X
U Hamburg	J. Haller	0.66	CMS		X	X	
U Göttingen	A. Quadt	0.5	ATLAS			X	
U Wuppertal	C. Zeitnitz	0.5	ATLAS	X			
Assoziiert							
GSI	K. Schwarz	-	ALICE	X	X	X	X
CERN	M. Elsing	-	ATLAS			X	X
DESY	V. Gülzow	-	verschiedene		X	X	X
GridKa	A. Petzold	-	verschiedene		X	X	X

Themenbereich 2: Milestones at Frankfurt

AP II-2

- M1: Erstellen einer vollständigen Liste der relevanten Unterschiede von XCache und “Disk Caching on the fly” und eines daraus abgeleiteten Entwicklungsplans (Q3/2022)
- M2: Vorhandensein eines gemeinsame Disk-Caching- Prototyps, der die Vorteile von XCache und “Disk Caching on the fly” vereint, vorzugsweise als Teil des Basiscodes von XRootD (Q3/2023)
- M3: Einbinden eines client-seitigen Caching-Systems für Lustre in die vorhandenen Job-Verteilungs- und Datenmanagementworkflows (Q2/2024)

AP II-4

- M1: Erster Prototyp eines Data-Lakes für FAIR, basierend auf den in Abschnitt 2.1.3 beschriebenen Vorarbeiten und Techniken (Q4/2022)
- M2: Einbau eines Hash-basierten Datenplatzierungs- und - Replikationsmechanismus in den Data-Lake-Prototyp (Q3/2023)
- M3: Demonstration eines effizienten Datenzugriff eines HPC- Zentrums auf den DataLake mit den entwickelten Cache- Technologien, Performance-Tests (Q2/2024)

Themenbereich 2: Frankfurt status

Decision for further investigation and developments based on XCache (with direct cache access)

Investigation/optimisation of possible delays when caching proxy is overloaded with a long queue of requests (solutions to circumvent/reduce overload of redirector)

Towards realistic workflows (also from other areas of FIDIUM)

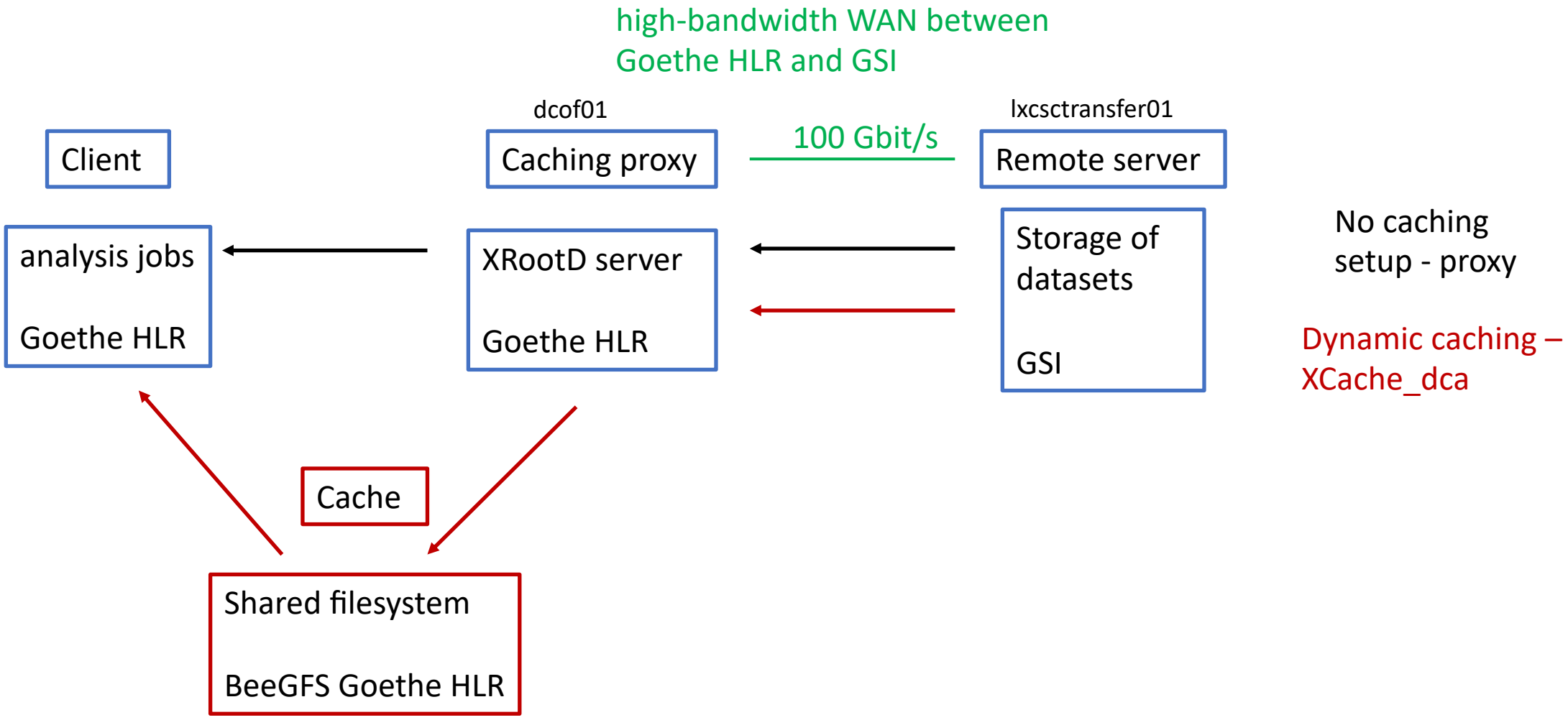
Set of use cases for efficient deployment of dynamic caching in terms of most relevant parameters

Coordination with other FIDIUM partners and XRootD/XCache developers foreseen

More details: → Presentation at FIDIUM annual meeting 2022

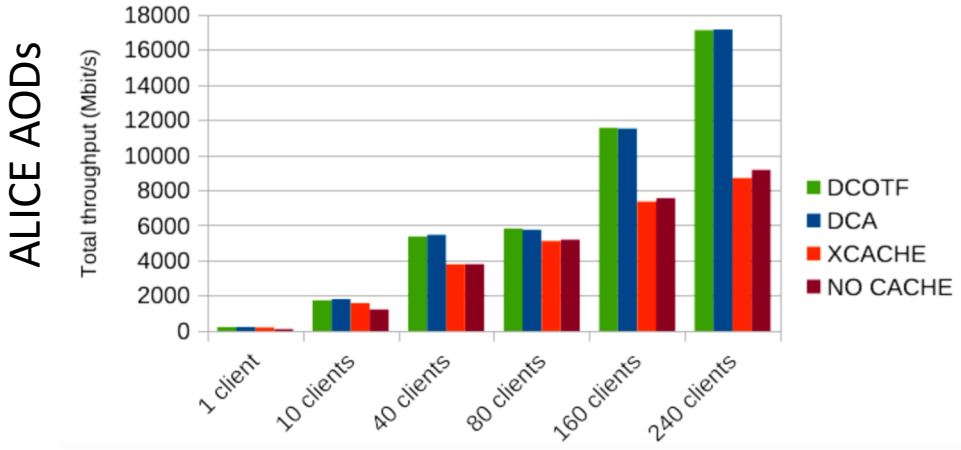
→ Some plans on following slides

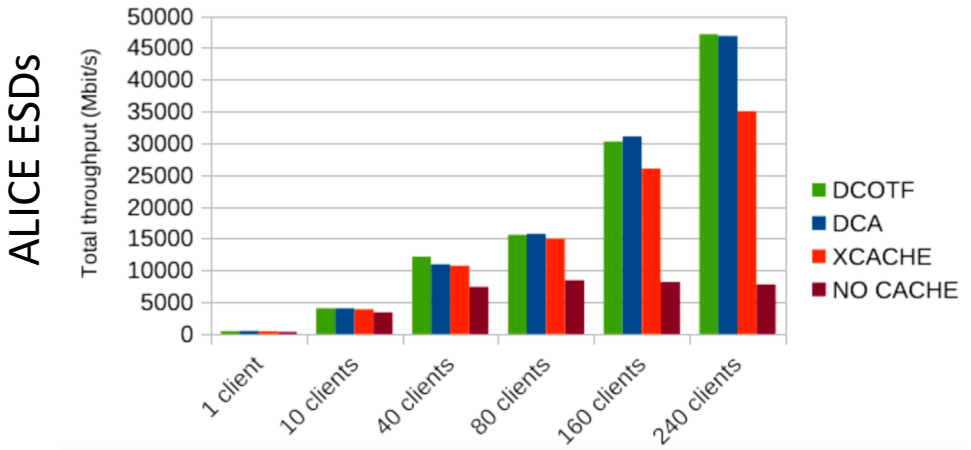
Network and setup: WAN Frankfurt- GSI



Discussion: Use of full bandwidth between Goethe HLR and GSI

Maximum measured bandwidth (in ErUM-IDT or FIDIUM) significantly below 100 Gbit/s





Limitations:

- Number and size of datasets at remote server (lxcsctransfer01)
- Access to remote server
- Processing in “deprecated“ analysis jobs
- Resources (number of nodes) at Goethe HLR

Documentation: Data lake prototype

Contents

SciTokens in PUNCH4NFDI: Update on January 17 (general meeting)

→ Related to milestone First prototype of a data lake for FAIR, based on technologies described in section 2.1.3

Discussion: Data lake prototype – requirements/developments

Previous work: Paul Kramp‘s master thesis

Requirements (documented also in “section 2.1.3“ of proposal):

- Local access point (connection to external clients)
- Standard protocols (e.g. http)
- Token-based authentication (application of SciTokens as implementation of OAuth)
- Global namespace in data lake (cluster/cache, using e.g. XRootD or Dynafed)
- Integrations of frameworks for mass storage e. g. dCache or XRootD
- Mechanisms for data placement and replication e. g. “Consistent Hashing” (distributed hash-tables)
- Optimisation of data access using local, dynamic caching systems

Planned projects:

Developments for efficient mechanisms for data placement and replication (e. g. hash-based)

Performance-Tests for efficient data access in relevant use cases/workflows

Themenbereich 3: Anpassung, Test und Optimierung auf Produktions- und Analyse-Umgebungen - FIDIUM overview

Arbeitspaket 1: Integration, Tests, Optimierung und Deployment der entwickelten Dienste

Arbeitspaket 2: Spezifische Anpassung der Dienste an komplexe Workflows und Nutzung spezieller Technologien für die Analyse wissenschaftlicher Daten

Arbeitspaket 3: Support

Standort	PI	FTE	Experiment	AP 1	AP 2	AP 3
RWTH Aachen	A. Schmidt / M. Erdmann A. Stahl	2	CMS / Einstein Teleskop	X	X	
KIT	G. Quast / A. Streit	0.66	CMS	X	X	X
KIT	R. Engel	0.5	Auger/IceCube/ Einstein Teleskop	X		X
Uni Mainz	F. Maas / A. Brinkmann	0.66	PANDA	X		
Uni Wuppertal	C. Zeitnitz	0.66	ATLAS	X		
GU Frankfurt	V. Lindenstruth	1	ALICE/CBM	X	X	
LMU München	T. Kuhr / G. Duckeck	1	Belle II / ATLAS	X	X	
U Freiburg	M. Schumacher	0.8	ATLAS	X		X
U Hamburg	J. Haller	0.66	CMS	X	X	
Uni Göttingen	A. Quadt	0.66	ATLAS	X		
Assoziiert						
GSI	K. Schwarz	-	ALICE	X		X
DESY	V. Gülzow	-	verschiedene	X	X	X
GridKa	A. Petzold	-	verschiedene	X	X	X

Themenbereich 3: Milestones at Frankfurt

AP III-1

- M1: Funktionale Tests (hinsichtlich Performanz und Skalierung) von Workflows mit dynamischem Caching (Q3/2022)
- M2: Funktionale Tests bei Workflows mit hohen Anforderungen bei Nutzung von GPU-Ressourcen zur Rekonstruktion von Experimentdaten (Q4/2023)

AP III-2

- M1: Implementierung zentraler Datenstrukturen repräsentativer Rekonstruktionsalgorithmen in vektorisierter Form, beispielsweise unter Verwendung der Klasse Vc (Q2/2023)
- M2: Laufzeitoptimierungen für die schnelle parallele Analyse großer Datenmengen durch Nutzung moderner Vektor- basierter Rekonstruktionsalgorithmen (Q2/2024)

Themenbereich 3: Frankfurt status/plans

Milestone: Funktionale Tests (hinsichtlich Performanz und Skalierung) von Workflows mit dynamischem Caching

Extended previous measurements for performances:
→ Statistics
→ Granularity in number of clients
→ **Number of AOD datasets**

Towards better coverage/understanding of multi-dimensional parameter space:
- Hardware resources for caching
- Datasets for analysis (total throughput higher for larger datasets)
- Numbers of clients and nodes for analysis
- Fraction of overhead for runtimes
- Network bandwidths

→ Scaling of runtimes can be optimized for different parameters
→ Basis of future integration into efficient workflows in data lakes

More details: → Presentation at FIDIUM annual meeting 2022
→ Talk by Grigory Kozlov today

Points for collaboration/discussion

Usage of “remote server“ lxcsctransfer01

Access rights and tools (Git, Slurm)

Organisation of regular meetings

Synergies: Developments at GSI/Frankfurt/PUNCH4NFDI

BACKUP

Connection to lxcsctransfer01

Copy from “remote server“ lxcsctransfer01 in October:

```
redelbach@dcof01:~$ xrdcp -f root://172.16.0.1:1094//data/AliAOD.root /tmp/test_AR/
[582.4MB/582.4MB][100%][==================================================][582.4MB/s]
```

Copy from “remote server“ lxcsctransfer01 yesterday:

```
redelbach@dcof01:~$ xrdcp -f root://172.16.0.1:1094//data/AliAOD.root /tmp/test_AR/
[0B/0B][100%][==================================================][0B/s]
Run: [FATAL] Connection error:  (source)
```

```
redelbach@dcof01:~$ ping 172.16.0.1
PING 172.16.0.1 (172.16.0.1) 56(84) bytes of data.
64 bytes from 172.16.0.1: icmp_seq=1 ttl=64 time=0.959 ms
64 bytes from 172.16.0.1: icmp_seq=2 ttl=64 time=1.05 ms
```

With GSI network:

```
aredelb@lxi097:~$ ssh lxcsctransfer01.gsi.de
aredelb@lxcsctransfer01.gsi.de's password:
Connection closed by 10.20.4.122 port 22
```