Föderierte Digitale Infrastrukturen für die Erforschung von Universum und Materie (FIDIUM)

Gemeinsamer Antrag von Gruppen aus den Bereichen Elementarteilchenphysik, Hadronen- und Kernphysik und Astroteilchenphysik

- Rheinisch-Westfälische Technische Hochschule Aachen, Prof. Dr. Alexander Schmidt[1]
- Rheinische Friedrich-Wilhelms-Universität Bonn, PD Dr. Philip Bechtle
- Goethe Universität Frankfurt am Main, Prof. Dr. Volker Lindenstruth
- Albert-Ludwigs-Universität Freiburg, Prof. Dr. Markus Schumacher
- Georg-August-Universität Göttingen, Prof. Dr. Arnulf Quadt
- Universität Hamburg, Prof. Dr. Johannes Haller
- Karlsruher Institut für Technologie, Prof. Dr. Günter Quast
- Johannes Gutenberg-Universität Mainz, Prof. Dr. Frank Maas
- Ludwig-Maximilians-Universität München, Prof. Dr. Thomas Kuhr
- Bergische Universität Wuppertal, Prof. Dr. Christian Zeitnitz

Assoziierte Partner sind

- CERN, Dr. Markus Elsing
- DESY, Prof. Dr. Volker Gülzow
- GridKa, Dr. Andreas Petzold
- GSI Helmholtzzentrum für Schwerionenforschung, Dr. Kilian Schwarz[2]

[1] Sprecher des Verbundes
[2] Stellvertretender Sprecher des Verbundes

1 Wissenschaftlicher Kontext

Die enormen wissenschaftlichen Fortschritte der experimentellen Teilchen-, Astroteilchen- sowie der Hadronen- und Kernphysik der letzten Jahrzehnte sind durch eine wachsende Bedeutung der Methoden und Ressourcen zur Datenauswertung gekennzeichnet. Mit steigender Präzision der Messinstrumente entstehen größere Datenmengen, die umfangreichere Rechenleistung und Speicherplatz erfordern. Darüber hinaus ist der nutzerfreundliche Zugang zu Computing- und Datenspeicher-Ressourcen von zunehmender Bedeutung.

Eine hervorragende Computing- und Massenspeicher-Infrastruktur ist heute zwingende Voraussetzung für exzellente Forschung.

Um die Exzellenz der deutschen Grundlagenforschung in den genannten Bereichen weiterhin zu sichern, sind in den kommenden Jahren erhebliche Anstrengungen erforderlich [1]. So wird z.B. im Bereich des LHC-Computings ab dem Jahr 2028 (Run 4) etwa ein Faktor 60 der Ressourcen von 2016 benötigt. Jedoch wird der technologische Fortschritt die Rechenleistung nur um einen Faktor 6 bis 10 erhöhen [2].

Auch die Experimente bei FAIR und den neuen Großprojekten der Astroteilchenphysik werden künftig ähnliche Anforderungen haben. Es werden daher experiment- und fachübergreifende Lösungen benötigt.

Eine zentrale Rolle bei Ansätzen zum Erreichen dieser Ziele werden die unter dem Begriff “Cloud Computing” subsumierten Technologien spielen. Für die Anwendungen in den genannten Communities am bedeutsamsten ist der Ansatz, komplette Infrastrukturen als Dienst zu implementieren (IaaS, Infrastructure as a Service). So werden nicht nur einzelne Applikationen oder Dienste, sondern komplexe Infrastrukturen aus einer Vielzahl von Diensten komplett virtuell bereitgestellt.

Die dabei zu nutzenden Ressourcen setzen sich aus Installationen mit dedizierten CPU- und Speichersystemen an großen Rechenzentren, Analysefarmen an Instituten und nur zeitweise verfügbaren Ressourcen an Partnerinstituten und HPC-Zentren (inkl. Supercomputing-Ressourcen der Gauß-Allianz) zusammen. Aber auch kommerzielle Cloud-Systeme sowie Cloud-Zugänge zu Ressourcen an Universitätsinstituten und Großforschungseinrichtungen sollen eingebunden werden. Auch die Nutzung von GPU-Clustern ist eine attraktive Option, die untersucht wird. Für den effektiven und robusten Betrieb der jeweiligen Experimentsoftware in diesem sehr heterogenen Umfeld ist ein hoher Grad der Abstraktion der Arbeitsabläufe und die Entwicklung entsprechender Softwarewerkzeuge notwendig. Nur so wird es möglich sein, eine Vielzahl an Standorten zu nutzen, ohne spezielle Fachexpertise aus der jeweiligen Nutzer-Community vor Ort verfügbar haben zu müssen. Durch die Übertragung derartiger Konzepte auf den Bereich der existierenden LHC Tier-1- und Tier-2-Zentren könnte auch der dort bisher notwendige Personaleinsatz reduziert werden.

Durch mehr föderierte und gemeinsam nutzbare Ressourcen können diese flexibler genutzt werden. Für unterschiedliche Workflows sollen die dafür optimalen Ressourcen ansteuerbar werden, so dass diese effizienter genutzt werden können (z.B. GPU-fähige Codes auf entsprechenden GPU-Ressourcen). Durch die damit verbundene Verbesserung der Auslastung kann ein Beitrag zur Schließung der Ressourcenlücke geleistet werden.

Während Rechenressourcen (CPU, GPU) flexibel und temporär nutzbar gemacht werden können (vgl. Abschnitt 3.1), sind die dabei zu prozessierenden Datensätze, die zukünftig in der Exabyte Größenordnung zu erwarten sind, weit weniger flexibel zu handhaben. Die langfristig angestrebte Lösung beinhaltet moderne Massenspeichersysteme, die auf wenige große Zentren konzentriert sind, z.B. am KIT, DESY, GSI, und die in der Form von sogenannten „Data Lakes“ zusammengefasst sind und einen einheitlichen Zugangspunkt anbieten. In der Kombination mit temporärem Daten-Caching an den jeweiligen heterogenen Computing-Ressourcen lässt sich daraus eine effiziente Infrastruktur aufbauen. Die Verwaltung lokaler Massendaten an kleinen Zentren, z.B. Universitätsgruppen, kann dadurch zum Teil reduziert bzw. wartungsfrei gestaltet werden und dafür entsprechende Daten-Caches automatisiert eingerichtet werden. Generell ist ein Betriebsmodell, in dem nur noch wenige große Zentren Massenspeichersysteme betreiben, preisgünstiger, weil entsprechend weniger Experten koordinierter und zentrierter eingesetzt werden können. Auch werden die Prozessabläufe durch eine Reduzierung der Anforderungen an die Softwareebenen oberhalb des Data-Lakes dadurch vereinfacht, dass die Ziele der Platzierung und der Replizierung von Datensätzen auf die Speicherebene verlagert werden sollen.

Durch intelligentes Daten-Management, z.B. durch optimierte Auslastung schneller (und damit teurer) Massenspeicher gegenüber langsamen (und preiswerten) Massenspeichern lassen sich erhebliche Kosteneinsparungen realisieren.

Dieser Antrag stellt eine Kontinuität in den Themenbereichen A und B der im Jahre 2018 ins Leben gerufenen Pilotmaßnahme ErUM-Data - “Innovative Digitale Technologien für die Erforschung von Universum und Materie” (IDT-UM Kollaboration) sicher. Eine entsprechende Fortsetzung ist notwendig, um im Jahre 2024 Technologien zur Hand zu haben und um bereits frühzeitig fundierte Entscheidungen für den HL-LHC treffen zu können [1]. Die beabsichtigte Laufzeit des Projektes ist 01.07.2021 bis 30.06.2024.

2 Stand der Technik

Viele in der Hochenergiephysik genutzte Werkzeuge unterstützen bereits moderne Cloud-Technologien. Darauf aufbauend soll das beantragte Projekt die Werkzeuge weiter entwickeln und für die deutsche Infrastruktur nutzbar machen. Ein bekanntes Beispiel ist CERNVM [3] und das dazu gehörige CERNVM-FS Dateisystem. Es handelt sich dabei um virtuelle Maschinen zum einfachen Einsatz auf mannigfaltigen Cloud-Ressourcen. CERNVM-FS ermöglicht die weltweite Softwareverteilung, was mittels Caching auf herkömmlichen HTTP-Proxyservern nahezu beliebig skalierbar ist. Fernzugänge zu experimentellen Daten werden durch das XRootD-Protokoll [4] ermöglicht.

Die Pilotmaßnahme ErUM-Data hat weitere Wege aufgezeigt, um den im Kapitel 1 erwähnten Herausforderungen zu begegnen. Durch das IDT-UM Verbundprojekt wurde eine Gruppen- und experimentübergreifende Zusammenarbeit ermöglicht, die es sonst so nicht gegeben hätte. Das Verbundprojekt ist außerdem in den Experimenten und international sichtbar. Es wurde z.B. auf dem Workshop von HEP Software Foundation (HSF), Open Science Grid (OSG) und WLCG neben den vergleichbaren Projekten in den USA (IRIS-HEP) und UK (IRIS)

präsentiert. Auch auf der International Conference on Computing in High Energy and Nuclear Physics (CHEP) gab es mehrere Vorträge der IDT-UM-Kollaboration. Wir sind Partner im Software Institute for Data Intensive Sciences (SIDIS), das sich zur Zeit im Aufbau befindet. Wissenschaftler des Verbundprojekts sind an der Entwicklung der WLCG-Strategie für Data Organization, Management, and Access (DOMA) beteiligt.

Demonstriert wurde zudem die dynamische virtuelle Erweiterung von Tier-1-Zentren prototypisch am Beispiel des GridKa am KIT. Hier werden temporär verfügbare Ressourcen an der Universität Bonn (Tier-3 und HPC), dem KIT (Tier-3 und HPC), sowie Cloud-Ressourcen am Leibniz-Rechenzentrum über ein Grid Compute Element (CE) experiment-übergreifend und transparent in das WLCG eingebunden.

Dabei hat sich herausgestellt, dass für eine weitere großskalige Integration von temporär verfügbaren Ressourcen die gesamte benötigte Infrastruktur, bestehend aus einer validierten Softwareumgebung, Datentransferdiensten, Datenbankzugriffen und lokalen Caches für Softwarepakete und Daten automatisiert und skalierbar aufgesetzt und mit einem Accounting- und Monitoringsystem ausgestattet werden muss.

Zusätzlich haben in Deutschland die ATLAS-Gruppe in Freiburg sowie die CMS- und Astroteilchenphysik-Gruppen in Karlsruhe diesbezügliche Erfahrungen beim Betrieb eines von diversen Anwendergruppen genutzten universitären HPC-Zentrums gesammelt. So wird in Freiburg zur Zeit ein vollständig virtualisiertes Tier-3 für das CMS-Experiment betrieben, das an ein in Karlsruhe betriebenes, CMS-typisches Workflowmanagementsystem angeschlossen ist. Auf dem selben Cluster wird von der Freiburger Gruppe ein voll virtualisiertes Tier-3 für ATLAS betrieben. Auch an der RWTH Aachen wird derzeit ein virtuelles Tier-2 Zentrum in kleinem Format im universitären HPC-Zentrum aufgesetzt, zusätzlich zum bereits existierenden dedizierten CMS Tier-2 Zentrum.

Eine lohnende Option ist die Nutzung von kurzzeitig nicht belegten HPC-Zentren, die wieder freigegeben werden, wenn höher priorisierte Nutzer die Ressourcen anfordern (sog. ”Backfilling”). Dieser Betriebsmodus wurde am KIT für einige Experimente bereits erfolgreich getestet und wird auch an HPC-Systemen am LRZ (SuperMUC-NG) regelmässig verwendet.

Eine Schlüsselrolle bei der dynamischen Nutzung heterogener Ressourcen spielt dabei die Anpassung, Nutzung und Weiterentwicklung der im Rahmen des Verbundprojektes geförderten Software `COBalD` (The opportunistic balancing daemon) [5] und `TARDIS` (Transparent Adaptive Resource Dynamic Integration System) [6].

Erste Erfolge bzgl. der Datencaches wurden ebenfalls erreicht. Datencaches müssen schnell und ohne großen Aufwand auch auf neu hinzugekommenen heterogenen Ressourcen realisiert werden und dann transparent in die vorhandene Produktions-Infrastruktur eingebunden werden.

Bei GSI ist ein auf XrootD basiertes “Disk Caching on the fly” genanntes System entwickelt worden, welches aktuell am Goethe-Hochleistungsrechenzentrum in Frankfurt und von der Freiburger Gruppe auf dem lokalen Tier2/3 für ATLAS aufgesetzt wird. KIT hat bereits vereinfachte Cache-Systeme am HPC-Zentrum bwForCluster NEMO in Freiburg sowie am KIT-Tier-3-Zentrum aufgesetzt. Es ist geplant, diese später auf das GSI-System umzustellen. Darüber hinaus hat KIT mit Eigenmitteln ein erweitertes XRootD-Monitoring entwickelt,

mit dem die effiziente Anwendbarkeit von Caching-Systemen auf verschiedene Workflows untersucht werden kann sowie Daten für eine Simulation des Verhaltens von verteilten Cache-Systemen gewonnen werden können. Die LMU München untersucht das Verhalten des ebenfalls XRootD-basierten Caching-Systems XCache im Produktionsbetrieb. Vorzugsweise kann hier eine Fusionierung unter Verwendung der besten Komponenten beider Projekte erfolgen.

Ein Beispiel für ein aktuell verwendetes System für föderiertes Datenmanagement ist das open-source Rucio-System, welches ursprünglich von ATLAS entwickelt wurde und nun von verschiedenen Communities weiterentwickelt und eingesetzt wird. Rucio beherrscht heute schon die Verwaltung von großen verteilten Datensätzen inklusive Quality-of-Service (QoS) Funktionalität und kommt auch in Data-Lake - Prototypen, z.B. im Rahmen des ESCAPE-Projekts, zum Einsatz.

Darüber hinaus haben die LMU München und die Universität Wuppertal user job monitoring-Systeme entwickelt, womit Anomalien gefunden werden können. Freiburg hat Entwicklungen zu `COBalD/TARDIS` in Form von Monitoring-Plugins und einem SLURM-Adapter zur Integration eines SLURM-Overlaybatchsystem beigetragen.

Des Weiteren stellen Wuppertal und GSI Workflows und Container-Images zur Verfügung, mit deren Hilfe Produktionsjobs in der Hochenergiephysik (Simulation, Rekonstruktion, Datenanalyse) ausgeführt werden können.

3 Förderziele und Arbeitsstruktur

Die Arbeitsziele sind in drei übergeordnete Themenbereiche gegliedert. Themenbereich I bezieht sich auf die Entwicklung von Werkzeugen und Technologien zur Einbindung und Nutzbarmachung heterogener Computing-Resourcen, während Themenbereich II die Einbindung von “Data-Lakes” und dazugehöriger Caches weiterentwickeln soll. Die Anpassung und Erprobung der in I und II entwickelten Technologien auf konkreten Zielsystemen sowie die Evaluation der Performance in kombinierten Tests wird in Themenbereich III behandelt. Die Themenbereiche sind jeweils in mehrere Arbeitspakete untergliedert. Die Beteiligung der einzelnen Verbundpartner sind in den jeweiligen Themenbereichen in Tabellen dargestellt.

3.1 Themenbereich I: Entwicklung von Werkzeugen zur Einbindung heterogener Ressourcen

Ziel des Themenbereichs I ist die ergebnisorientierte, gemeinschaftliche und fachübergreifende Entwicklung von Werkzeugen und Strukturen zur Einbindung und effizienten Nutzung von heterogenen Ressourcen. Die beteiligten Arbeitsgruppen werden dabei insgesamt an zwei Arbeitspaketen arbeiten, welche im folgenden genauer beschrieben sind.

3.1.1 Arbeitspakete

Aus den oben genannten Anforderungen ergeben sich die folgenden Arbeitspakete:

1. Erschließung und effiziente Einbindung von opportunistischen Ressourcen
    - Weiterentwicklung und Anpassung des Resource Managers `COBalD/TARDIS` an zukünftige Gegebenheiten
    - Entwicklung von dynamischer Steuerung des Job-Schedulings (z.B. Berücksichtigung von Datenlokalität, I/O-Raten)
    - Automatisierte Skalierung peripherer Dienste
    - „Compute Site in a Box“: Nutzbarmachung der Ressourcen an Tier-3- und Tier-2-Zentren mit minimalen zusätzlichen administrativen Ressourcen (volle Automatisierung, Skalierbarkeit)
2. Accounting und Controlling von heterogenen Ressourcen
    - Werkzeuge zum Accounting der opportunistisch genutzten Ressourcen
    - Tools für kontinuierliche Überwachung der Nutzungseffizienz

Arbeitspaket 1: Erschließung und effiziente Einbindung von opportunistischen Ressourcen: Die Einbindung von geeigneten opportunistischen Rechenressourcen an HPC-Zentren (CPU und GPU), lokalen Institutsclustern bzw. bei wissenschaftlichen oder privaten Cloud-Anbietern in die existierenden spezifischen Infrastrukturen des wissenschaftlichen Rechnens gewinnt für die beantragenden Nutzergruppen mehr und mehr an Bedeutung. Auf Grund der durch die Vielfalt der zur Verfügung stehenden Ressourcen zunehmenden Heterogenität und der verschiedenen Anforderungen der Nutzergruppen ist eine sinnvolle Integration nur mittels Verwendung von Container- oder Virtualisierungstechnologien möglich. Gleichzeitig gilt es, die Komplexität der heterogenen Ressourcen vom Anwender durch eine dynamische und transparente Integration in die im Rahmen der Pilotmaßnahme ErUM-Data prototypisch konzipierten einheitlichen Infrastruktur mit einem wohldefinierten Zugangspunkt zu kapseln, um so eine einfache und effiziente Nutzung der Ressourcen zu gewährleisten. Hierzu soll der in der Pilotmaßnahme entwickelte „Resource Manager“ `COBalD/TARDIS` verwendet und entsprechend der zukünftigen Gegebenheiten weiterentwickelt werden. Analog dazu bedarf es noch weiterer Entwicklungen im Bereich der oben genannten einheitlichen Infrastruktur. So ist unter anderem die Berücksichtigung von Datenlokalität beim Scheduling der Jobs in Verbindung mit Themenbereich II sowie die Steuerung von Workflows auf geeignete Ressourcen anhand von den Experimenten zur Verfügung gestellten charakteristisch beschreibenden Parametern (z. B. I/O-Raten etc.) sicherzustellen. Auch im Bereich der peripheren Services (z. B. den lokalen Web-Caches für Datenbankzugriffe und CVMFS) sind Entwicklungsarbeiten notwendig, um eine automatisierte Skalierung dieser Services mit der dynamisch schwankenden Anzahl von Ressourcen zu ermöglichen.

Um die Einstiegshürden auf Seiten der Anbieter von opportunistischen Ressourcen zu reduzieren und damit das Angebot solcher Ressourcen weiter zu erhöhen, soll das „Compute Site in a Box“ Konzept implementiert werden: Ziel ist die Entwicklung einer mit minimalen zusätzlichen Ressourcen einsetzbaren Lösung, um Tier-3-Zentren für zentral organisierte Rechenaufgaben der Großexperimente nutzbar zu machen. Die opportunistische Nutzung

der vom Hauptnutzer zeitweise nicht benötigten oder nicht nutzbaren verteilten Compute-Ressourcen bietet ein enormes Potenzial zur Effizienzsteigerung von HPC- und HTC-Sites, gleichzeitig müssen jedoch die zusätzlich nötigen Personalressourcen minimiert werden.

Teil der „Compute Site in a Box“ ist die Automatisierung aller Dienste, die für den Zusammenschluss der verfügbaren Sites in einer Cloud benötigt werden. Dazu können je nach Gegebenheiten des Standorts z.B. reine Rechenknoten, lokale Data Caches oder die zum Betrieb benötigten Dienste wie hochverfügbare Web-Caches oder ein „Resource Manager“ gehören. Um eine optimale Nutzung der Ressourcen zu ermöglichen, ohne dabei signifikante zusätzliche Administrationsaufgaben sowohl in den dezentralen Tier-3s als auch im zentralen Management der Experimente zu erzeugen, bietet diese Lösung eine vollständige Automatisierung und einfache Skalierbarkeit. Auch für universitäre Tier-2-Zentren, an denen die personelle Situation oft recht angespannt ist, wäre der Einsatz einer solchen Lösung attraktiv. Die Automatisierung muss dabei modular, möglichst portabel und ohne Community-Spezifika strukturiert sein und gleichzeitig Anpassungen für Spezifika der einzelnen Standorte erlauben. Die Federführung dieses Arbeitspaketes obliegt der Universität Bonn und dem KIT.

Arbeitspaket 2: Accounting und Controlling von heterogenen Ressourcen: Ein weiteres Hauptaugenmerk des Themenbereichs I stellt die Entwicklung von Werkzeugen zur präzisen Nachverfolgung und Dokumentation der reservierten und tatsächlich verwendeten Ressourcen dar. Dies dient auch der Überwachung der Leistungsfähigkeit der eingebundenen opportunistischen Ressourcen.

Da gerade im Kontext der Verwendung von verschiedenen heterogenen Ressourcen (wie z.B. HPC-Cluster, HTC-Cluster, Universitäts-Cluster) diverse Informationen der beteiligten Komponenten (Overlay-, Backend-Batchsystem und Metaschedulers) zusammengeführt werden müssen, um eine korrekte Abrechnung der verwendeten Ressourcen gewährleisten zu können, stellt diese Aufgabe eine große Herausforderung dar. Es soll ein umfassendes Accounting-System bereitgestellt werden, das die Hauptroutinen geeignet abstrahiert, sodass die verwendeten Technologien der darunterliegenden Systeme (HTCondor, `COBalD`/`TARDIS`, OpenStack, SLURM,…) einfach durch diverse Plugins angepasst bzw. neue Systeme hinzugefügt werden können.

Für die Anbieter ist eine genaue Abrechnung der durch die beantragenden Nutzergruppen opportunistisch genutzten Ressourcen unerlässlich, um einerseits die Verwendung gegenüber den Stakeholdern nachzuweisen und andererseits um die Bereitstellung der Ressourcen vom jeweiligen Experiment entsprechend akkreditiert zu bekommen. Außerdem können die zusätzlichen Informationen der tatsächlich abgearbeiteten Operationen in ein intelligentes Workflowmanagement integriert werden. Dafür ist eine kontinuierliche Überwachung der eingebundenen opportunistischen Ressourcen zwingend erforderlich, z.B. um eine möglichst effiziente Nutzung der Ressourcen sicherzustellen und im Bedarfsfall automatisiert eingreifen zu können. Die Federführung dieses Arbeitspaketes obliegt der Universität Freiburg.

3.1.2 Beteiligte Institute, Koordination, beantragte Mittel

Die beteiligten Partner im Themenbereich, sowie die beantragten Mittel sind in Tabelle 1 aufgeführt. Die Koordination des Themenbereichs wird übernommen von Dr. Manuel Giffels und Dr. Oliver Freyermuth.

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

Tabelle 1: Standorte, beantragte Mittel und assoziierte Partner im Themenbereich I. Eine Tabelle mit der Gesamtübersicht aller Themenbereiche findet sich am Ende des Dokuments.

Expertise: Die beteiligten Partner vom KIT, der Uni Bonn, der Uni Freiburg, der Uni Frankfurt am Main, der Uni Göttingen, der Uni Wuppertal, dem GSI und DESY bringen große Expertise im Betrieb von Rechenzentren im Kontext der Kern- und Teilchenphysik (WLCG) mit. Der maßgeblich am KIT entwickelte und mit den Partnern in IDT-UM weiterentwickelte Ressourcen-Manager `COBalD/TARDIS` wird bereits erfolgreich am KIT, in Bonn, Freiburg und an der LMU zur opportunistischen Nutzung lokaler HPC/HTC-Zentren in einer gemeinsamen Infrastruktur mit einheitlichem Einstiegspunkt am KIT eingesetzt und zum Teil mitentwickelt. Zur vollen Automatisierung aller Tier-3-Dienste inklusive `COBalD/TARDIS` liegen in Bonn langjährige Erfahrungen vor, die als „Compute Site in a Box“ verallgemeinert und mit den Partnern weiterentwickelt werden sollen. Freiburg kann durch die Entwicklung und Koordination des experimentübergreifenden *HammerCloud*-Projekts jahrelange Erfahrungen beim Testen und Monitoren von Compute-Sites einbringen und hat als wichtige Grundlage für präzises Accounting bereits dedizierte Monitoring-Plugins für `COBalD/TARDIS` entwickelt. Die Universität Frankfurt hat umfangreiche Erfahrung in der effizienten Nutzung heterogener GPU/CPU Ressourcen und im Bereich der architektur-übergreifenden Aufteilung von Prozessierungsschritten, sowohl am Goethe-HLR-Cluster als auch bei der Entwicklung des ALICE High Level Triggers.

3.2 Themenbereich II: Data-Lakes, Distributed Data, Caching

Ziel des Themenbereichs II ist die ergebnisorientierte, gemeinschaftliche und fachübergreifende Entwicklung von Technologien für den Aufbau von effizienten und intelligenten föderierten Datenspeicherinfrastrukturen. Alle Entwicklungen werden in den internationalen Kontext eingebunden sein und die Bedarfe der großen Experimentkollaborationen berücksichtigen. Die beteiligten Arbeitsgruppen werden dabei insgesamt an vier Arbeitspaketen arbeiten, die im folgenden genauer beschrieben sind. Im ersten Arbeitspaket soll ein Echtzeit-Monitoring-System für Data-Lakes entwickelt werden. Die dynamischen Komponenten der folgenden Arbeitspakete werden auf den Informationen aus dem Monitoring-System aufbauen. Im zweiten Arbeitspaket sollen dynamische Daten-Caches untersucht werden. Im dritten Arbeitspaket werden intelligente Datenplazierungs und -Replikationsmechanismen entwickelt. Im vierten Arbeitspaket werden prototypische Data-lakes für die beantragenden Communities aufgebaut und die effiziente Anbindung von Nutzern, Datenquellen und Compute-Infrastrukturen untersucht. Die dafür notwendigen Technologien müssen zum Teil entwickelt bzw. weiterentwickelt werden. Einige wichtige Vorarbeiten sind bereits in IDT-UM erfolgt.

3.2.1 Arbeitspakete

Die Arbeitspakete, ihre Titel und die zu bearbeitenden Themen werden im folgenden noch einmal aufgelistet. Anschließend wird jedes Arbeitspaket inklusive der von den beteiligten Partnern geplanten Arbeiten im Detail beschrieben:

1. Aufbau eines Echtzeit Data-Lake-Monitoring-Systems
    - Erfassung der Auslastung von Data-Lake-Komponenten
    - Erfassung von Datenzugriffsmustern
2. Technologien für Data-Lake-Caching
    - Weiterentwicklung und Konsolidierung von Daten-Cache - Technologien
    - effiziente Einbindung von dynamischen Datencaches in den Data-Lake und an CPU-Ressourcen
    - Einsatz von parallelen ad-hoc Filesystemen als Caches in HPC-Systemen
3. Technologien für Data-Lake-Daten- und Workflow-Management
    - Replikations- und Platzierungsmechanismen
    - bedarfsgetriebene Datenmanagement-Mechanismen
    - effizienter Datenzugriff und Anpassung an Workload-Management-Systeme
4. Data-Lake-Prototypen, Technologien für QoS und effiziente Anbindung
    - Aufbau von Data-Lake-Prototypen
    - effiziente Anbindung von Nutzern, Zentren und Datenquellen
    - Quality of Service

Arbeitspaket 1: Aufbau eines Echtzeit Data-Lake - Monitoring-Systems: Das Monitoring-System muss in der Lage sein, in Echtzeit die Auslastung und den Füllstand von einzelnen Speicherkomponenten im Data-Lake zu überwachen und die Platzierung von Datensätzen daran auszurichten, sowie oft und wenig oft verwendete Datensätze erkennen zu können und ebenso, von wo aus (auch in Bezug auf verfügbare Netzwerkbandbreite) häufig auf Datensätze zugegriffen wird. Basierend auf diesen Informationen muss der Data–Lake in der Lage sein, Datensätze zum Beispiel auf Bandspeicher auszulagern oder Daten entsprechend zur Zugriffsoptimierung zu kopieren oder zu replizieren.

In einem ersten Schritt sollen die nativ vom in den ErUM-Communities weit verbreiteten XRootD-Framework mitgelieferten Monitoring-Möglichkeiten implementiert, angepasst, evaluiert und gegebenenfalls erweitert werden. XRootD-Dienste bieten die Möglichkeit, Monitoring-Informationen als UDP-Pakete an dedizierte Kollektor-Maschinen zu senden, wo sie ausgewertet und dann entsprechende Aktionen daraus abgeleitet werden können. Die Informationen werden überdies im XML-Format geliefert und sind daher universell interpretierbar. Die hierdurch zur Verfügung stehenden Monitoring-Daten sollen dann zusammengefasst und über ein Web-Front-End den Administratoren der Systeme verfügbar gemacht werden.

Nach einer ausführlichen Bedarfs- und Gap-Analyse soll in einem zweiten Schritt das XRootD - basierte Data-Lake - Echtzeitmonitoringsystem ergebnisabhängig auf andere im Data-Lake verwendete Datenspeichersysteme, sowie auf die Bedarfe der einzelnen Communities wie z.B. der Astroteilchenphysik, und entsprechende Protokolle erweitert werden.

Die Arbeiten hierfür erfolgen in erster Linie an der Universität Wuppertal, bei GSI und am KIT (Astroteilchenphysik).

Arbeitspaket 2: Technologien für Data-Lake - Caching: Die Datenmengen in ErUM stellen eine immense Herausforderung hinsichtlich effizienter Methoden der Datenanalyse und -Verteilung dar. Das wesentliche Problem besteht in der mangelnden Datenlokalität, d.h. der Vorhaltung der Daten auf den durch die Rechnerressourcen nutzbaren Massenspeichern. Daher sollen Datencache-Technologien für heterogene Ressourcen (weiter) entwickelt werden. Die neuen Techniken müssen es ermöglichen, effizient funktionierende Datencaches schnell und auf einfache Weise und ohne großen Aufwand auch auf neu hinzugekommenen Ressourcen (z.B. HPC- und Cloud-Systemen) zu realisieren und selbige dann transparent in die vorhandenen Produktionsinfrastrukturen einzubinden.

Hierzu sollen die Vorarbeiten aus der Pilotmaßnahme ErUM-Data zum Einsatz kommen. Basierend auf den bei GSI entwickelten XRootD - Plug-Ins wurde ein Disk-Caching - Prototyp „Disk Caching on the fly“ entwickelt [7]. In intensiven Performance- und Vergleichstests wurden die Gemeinsamkeiten und Unterschiede zum bereits in XRootD integrierten Caching - System „XCache“ herausgearbeitet. In enger Zusammenarbeit mit dem XRootD-Entwicklerteam sollen nun die Vorteile beider Caching-Systeme zu einem gemeinsamen System fusioniert werden. Hierbei werden GSI, die Goethe-Universität Frankfurt, das KIT sowie die LMU München intensiven Erfahrungsaustausch betreiben. Bei der Weiterentwicklung von innovativen Caching-Technologien wird sich auch die Universtät Hamburg beteiligen sowie die JGU Mainz Erfahrungen bei der Integration des Hierarchical Storage Managements (HSM) von Lustre für

den Aufbau von Client-Side Caching-Systemen einbringen.

Anschließend sollen die entwickelten dynamischen Disk-Caching - Systeme für die Verwendung in heterogenen Ressourcen wie HPC- und Cloud-Systemen optimiert und weiterentwickelt sowie auch an die Bedarfe der einzelnen Experimente angepasst werden. Hierzu plant KIT eine Simulation von komplexen Systemen mit dynamischen Caches zur Optimierung der Hardware und Caching-Strategien. In diesem Zusammenhang soll auch ein bei KIT entwickeltes Konzept für koordiniertes verteiltes Caching evaluiert werden inklusive einer möglichen Einbindung in existierende Workload- und Datenmanagementsysteme. Speziell sollen gemeinsam mit der LMU München wichtige Themen, wie die Berücksichtigung der Caching-Inhalte bei der Job-Verteilung, untersucht werden. Die LMU München plant überdies die Optimierung der bei der Analyse verwendeten Datenformate für block-wise Caching zu untersuchen, sowie Konzepte für Cache-pre-loading zu erarbeiten. In die Optimierung des Gesamtsystems wird die JGU Mainz Verfahren zum Bandbreitenmanagement von parallelen Dateisystemen einbringen, so dass Stage-In Prozesse sowie die Nutzung von Remote-Lustre - Systemen nicht mit der Nutzung lokaler HPC-Anwendungen bzgl. der QoS Anforderungen kollidieren

Ein weiterer Ansatz ist die Verwendung paralleler Filesysteme, die mittels der Computeknoten selbst und den lokalen Speichersystemen der Computeknoten aufgespannt werden, sogenannte on-demand oder ad-hoc parallele Filesysteme. Parallele Dateisysteme als sogenannte „Burst Caches“ sind Gegenstand aktueller Forschung. Hier haben Mainz und KIT bereits gemeinsame Vorarbeiten im Rahmen des ADA-FS Projektes als Teil vom DFG-Schwerpunktprogramm SPPEXA gemacht. Insbesondere auf HPC-Systemen verfügen die meisten Knoten über lokalen SSD-Plattenplatz, welcher ungenutzt für die meisten parallelen Anwendungen brach liegt. Durch vom Nutzer ohne administrative Rechte erzeugbare parallele Dateisysteme, wie z.B. das in Mainz entwickelte Gekko-FS, können diese einen erheblichen Mehrwert bilden. Die Daten und Metadaten der temporären Caches können dabei über ein hierarchisches Speichermanagement mit den parallelen Dateisystemen der HPC Zentren konsistent gehalten werden.

Arbeitspaket 3: Technologien für Data-Lake - Daten- und Workflow - Management: Die zu entwickelnden dynamischen Data-Lake - Workflows sollen in der Lage sein, basierend auf den in Arbeitspaket 1 gesammelten Monitoring-Informationen zu agieren. Durch intelligentes Daten-Management und somit eine effiziente Ressourcennutzung lassen sich erheblich Kosten sparen. Bestehende Technologien, die bereits heute Anwendung finden, sollen getestet werden. Eines der zu evaluierenden Systeme wird das in Kapitel 2 beschriebene Rucio-System sein, welches viele der angedachten Funktionalitäten bereits zur Verfügung stellt. In diesem Zusammenhang soll auch die Verwendung eines experimentübergreifenden und einheitlichen Systems von Metadaten von KIT untersucht werden.

Parallel dazu sollen neuartige alternative Ansätze erforscht werden, wie z.B. eine „Hash“-basierte Datenplatzierung und -replizierung, wodurch man versuchen könnte, soweit wie möglich auf klassische Filekataloge zu verzichten. Eine mögliche Implementierung basiert auf „Consistent Hashing“. Weitere themenbezogene Details können mit Konzepten wie Regionen

und Zonen addressiert werden, wie das auch viele kommerzielle Cloud-Anbieter machen. Die hier beschriebenen Entwicklungsarbeiten werden vor allem am KIT, am CERN sowie an der GSI und an der Universität Göttingen erfolgen.

Basierend auf den beschriebenen Techniken sollen erste Prototypen für einen entsprechenden Datenplatzierungs- und Daten-Replikationsdienst für einen Data-Lake entwickelt werden.

Ausführliche Tests der beschriebenen Systeme werden am KIT erfolgen inklusive einer Evaluierung und eventuellen Anpassung für die Astroteilchenphysik.

Ein wichtiger Aspekt bei der effizienten Nutzung heterogener Ressourcen für die Analyse ist der schnelle Zugriff auf die zu prozessierenden Daten. In diesem Zusammenhang ist es geplant, lokal eingesetzte Workload-Manager derart zu erweitern, dass die Einstellung eines parallelen Filesystems on-demand bei Verwendung als Cache in HPC-Systemen, integriert werden kann. Gerade bei der Datenanalyse lässt sich nicht vorhersagen, welche Datensätze in Zukunft an welchen Standorten benötigt werden. Ein Schwerpunkt der Untersuchung wäre hierbei die effiziente automatisierte Auswahl der zu cachenden Daten, um für den Nutzer transparent eine hohe Datenleserate zu erreichen. Hierfür können bereits existierende Forschungsergebnisse von ATLAS und der LMU München sowie der JGU Mainz als Ausgangsbasis verwendet werden. Da die Analysejobs eines Nutzers immer stoßweise submittiert und damit häufig auch beendet werden, kann eine sehr große Anzahl gleichzeitiger Schreibversuche anfallen, die einzelne, entfernte Speichersysteme überlasten kann. Ein ausfallsicheres Zurückschreiben der Ergebnisdaten muss daher stets sichergestellt werden. Mit diesen Themengebieten werden sich die Universität Hamburg und die Universität Mainz beschäftigen.

Arbeitspaket 4: Data-Lake - Prototypen, Technologien für QoS und effiziente Anbindung: Im Rahmen dieses Arbeitspakets sollen Data-Lake - Prototypen für die ErUM-Wissenschaften zur Verfügung gestellt sowie Technologien für QoS und effiziente Anbindung von Nutzern, Datenquellen und Recheninfrastrukturen entwickelt werden. Die Erarbeitung der Data-Lake Konzepte soll in enger Zusammenarbeit und Abstimmung mit den Experimenten und den Data-Lake Konzepten aus WLCG/DOMA sowie nationalen und internationalen Initiativen wie der NFDI, EOSC und ESCAPE erfolgen.

Ein Data-Lake als zugrundeliegende Speicherinfrastruktur einer wissenschaftlichen Compute-Cloud sollte von Nutzern und Applikationen aus über einen einheitlichen Einstiegspunkt erreichbar sein, wobei der Ort, von dem aus der Client-Zugriff erfolgt, berücksichtigt werden sollte. Die Nutzerschnittstelle muss einfach bedienbar sein, sowohl via Kommandozeile als auch über eine grafische Benutzerschnittstelle und eine standardisierte Programmierschnittstelle muss ebenfalls zur Verfügung gestellt werden. Vorzugsweise kommunizieren Clients mit dem Data-Lake mit industriekompatiblen Standardprototollen, wie z.B. „http“. Der Zugriff auf den Data-Lake sollte überdies für Anwender transparent sein. Es ist von Vorteil für die globalen Workflow- und Datenmanagement-Systeme standardisierte Schnittstellen zur Verfügung zu stellen, wodurch auch experimentübergreifende Analysen unterstützt werden. Alle Daten müssen von überall aus über einen einheitlichen Namensraum erreichbar sein und von jedem Ort

des Data-Lakes aus sollte der vollständige Namensraum bekannt sein. Zur Authentifizierung sollten moderne Token-basierte Technologien zur Anwendung kommen. Letztendlich sollte der Data-Lake modular aufgebaut werden, so dass Komponenten und Technologien auch zu späteren Zeitpunkten noch durch neu aufkommende und eventuell besser geeignete Module ersetzt werden können. Ein einheitlicher Namensraum kann z.B. über Rucio, einen XRootD - basierten globalen Redirector, oder über Dynafed realisiert werden. Zur Optimierung von Datenzugriffen innerhalb des Data-Lakes aber auch bezogen auf Compute-Ressourcen, die an den Data-Lake angeschlossen werden, sollen lokale Cache-Systeme zum Einsatz kommen. Hier sind bereits intensive Entwicklungsarbeiten im Pilotprojekt erfolgt (vgl. Abschnitt 2).

Anhand der CAP-Eigenschaften (Consistency, Availability, Partitioning) können die QoS-Eigenschaften des Data-Lakes aufgezeigt und über intelligente QoS-Algorithmen versucht werden, die durch das CAP-Theorem gegebenen Einschränkungen zu reduzieren.

Alle existierenden Datenquellen, was auch die Daten aus Observatorien der Astroteilchenphysik einschließt, müssen effizient über offene Protokolle in den Data-Lake integriert werden. Auch müssen bestehende Computing-Systeme auf einfache Weise in die aufgebauten Data-Lake - Prototypen eingebunden werden können, wozu auch das global verteilte WLCG-Computing- System zählt. Daher sollen Techniken entwickelt werden, um alle existierenden Rechenressourcen, sowohl dedizierte experimenteigene Ressourcen als auch opportunistische Ressourcen, HPC- und Cloud-Systeme sowie dedizierte Analysezentren, performant an den Data-Lake anzubinden. Nutzer von all diesen Zentren aus müssen auf einfache Weise auf Analysedaten im Data-Lake zugreifen können.

Die hier beschriebenen Arbeiten werden von GSI, Goethe - Universität Frankfurt, KIT & GridKa, LMU München, Universität Mainz und CERN adressiert werden.

3.2.2 Beteiligte Institute, Koordination, beantragte Mittel

Die beteiligten Partner im Themenbereich, sowie die beantragten Mittel sind in Tabelle 2 aufgeführt. Die Koordination des Themenbereichs wird übernommen von Dr. Kilian Schwarz und Prof. Dr. Andre Brinkmann.

Expertise: Die beteiligten Partner aus dem Karlsruher Institut für Technologie (KIT), der Goethe-Universität Frankfurt am Main, der Johannes Gutenberg-Universität Mainz, der Ludwig-Maximilians - Universität München, der Universität Hamburg, der Georg-August - Universität Göttingen, dem GSI Helmholtzzentrum für Schwerionenforschung, DESY und dem CERN, haben als verantwortliche Gruppen für WLCG-Zentren alle langjährige Erfahrung im Betrieb von Rechenzentren für die Hochenergiephysik und teilweise bereits Expertise mit der opportunistischen Nutzung von heterogenen Ressourcen. GSI, die Universität Frankfurt, KIT und die LMU haben über IDT-UM bereits einschlägige Erfahrung in der Entwicklung und Optimierung von dynamischen Daten-Caches. GSI und CERN haben als Partner des EU-Projekts ESCAPE Erfahrung in der Entwicklung von Data-Lake - Technologien. KIT, DESY, GSI und CERN haben als Betreiber von zentralen Infrastrukturen überdies einschlägige Erfahrung mit dem Aufsetzen und Betreiben von Monitoring-Systemen. Auch die Arbeitsgruppe in Wuppertal

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

Tabelle 2: Standorte, beantragte Mittel und assoziierte Partner im Themenbereich II. Eine Tabelle mit der Gesamtübersicht aller Themenbereiche findet sich am Ende des Dokuments.

beschäftigt sich seit vielen Jahren mit Überwachungssystemen und entwickelt und wartet das ATLAS-weit verwendete LOCALGroup-Disk Monitoring.

3.3 Themenbereich III: Anpassung, Test und Optimierung auf Produktions- und Analyse-Umgebungen

Der erfolgreiche Test aller im Themenbereich I und II entwickelten Komponenten und deren Integration in ein Gesamtsystem ist die Voraussetzung, dass das entwickelte System auch längerfristig erfolgreich für den Betrieb heterogener Ressourcen zum Einsatz kommen wird und auch eine größere weltweite Verbreitung (z.B. WLCG-Zentren) haben wird. Hierfür müssen die entwickelten Werkzeuge zeigen, dass ein schlanker, robuster und effektiver Betrieb möglich ist. Weiterhin soll die Skalierbarkeit und das Zusammenspiel mit den unterschiedlichen Anforderungen von Forschungsprojekten untersucht werden.

Die Dienste und Techniken, die diese Anforderungen erfüllen, sollen Performance-, Skalierungs- und Robustheitstests unterzogen werden, um Schwachstellen aufzudecken und diese beseitigen zu können. Dies erfordert produktionsnahe Testumgebungen, wie sie z.B. an den WLCG-Tier-Zentren und Analyse-Zentren zur Verfügung stehen.

3.3.1 Arbeitspakete

Aus den oben genannten Anforderungen, ergeben sich die folgenden Arbeitspakete, anschließend wird jedes Arbeitspaket inklusive der von den beteiligten Partnern geplanten Arbeiten im Detail beschrieben.

1. Integration, Tests, Optimierung und Deployment der entwickelten Dienste
    - Integration der verschiedenen Komponenten: Workflowmanagement, Caching, Accounting, Ressourcenmanagement und Überwachungsysteme
    - Funktionale Tests auf ausgewählten Zentren
    - Integration in die Produktionsumgebung der beteiligten Experimente
    - Am Ende, nach erfolgreichen Tests, “Deployment” der Gesamtlösung an den verfügbaren WLCG-Tier-Zentren, HPC-Zentren und Cloud Anbietern.
2. Spezifische Anpassung der Dienste an komplexe Workflows und Nutzung spezieller Technologien für die Analyse wissenschaftlicher Daten
    - Optimierung für spezielle Workflows mit hoher IO Last, Speicherbedarf, GPU Nutzung, u.a.
    - Optimierung für schnelle parallele Analyse großer Datenmengen mit modernen vektor-basierten Analysealgorithmen
3. Support
    - Einrichtung eines standortübergreifenden Support-Teams, das die Zentren bei Installation und Betrieb unterstützt

Arbeitspaket 1: Integration, Tests, Optimierung und Deployment der entwickelten Dienste Die Bündelung der entwickelten Dienste muss einen effektiven und schlanken Betrieb der Dienste ermöglichen, inklusive eines ausgefeilten Überwachungssystems und der Erfassung der Nutzung (Accounting). Dies bedeutet, dass nach der Integration der verschiedenen Komponenten, wie Workflowmanagement, Caching-Lösungen, Accounting, Ressourcenmanagement und erweiterten Überwachungsystemen, funktionale Tests unter realistischen Bedingungen durchgeführt werden, die in repräsentative Anwendungen der beteiligten Partner münden. Hierfür sollen Installationen auf den im Verbund direkt zugänglichen Ressourcen genutzt werden (z.B. Tier-3 Zentren, Analyse-Cluster). Details zu den verfügbaren Ressourcen sind in Tabelle 4 angegeben.
In nächsten Schritt sind Tests im Produktionsumfeld der beteiligten Experimente geplant. Hier bringen die Partner des Projekts sehr viel Erfahrung mit und haben Zugriff auf die entsprechenden Computing-Systeme (z.B. WLCG-Tier-Zentren).
Nach erfolgreicher Integration und Tests, soll das “Deployment” der Gesamtlösung auf verschiedenen Systemen durchgeführt werden. Hierbei ist eine Zusammenarbeit mit WLCG-Tier-Zentren, HPC-Zentren (z.B. der Gauß-Allianz) und Cloud Anbietern geplant, aber auch eine

Erweiterung außerhalb von Deutschland ist vorgesehen.
Bei den genannten Tests sollen die typischen Anwendungsfälle (Simulation, Daten-Rekonstruktion und Analyse) abgedeckt werden. Dies macht die Integration der Cache bzw Workflow-Dienste in die Daten- bzw Workflow-Management Systeme der Experimente, in Zusammenarbeit mit WLCG und den Experimenten, notwendig. Hierbei soll die Zuverlässigkeit, Skalierbarkeit und Wartbarkeit der entwickelten Lösungen unter realen Bedingungen untersucht werden. Dies impliziert größere Tests an den Computing- und Analyse-Zentren aller beteiligter Partner, sowie an Cloud-Systemen anderer Anbieter (z.B. EU Open Science Cloud, kommerzielle Anbieter). Schließlich soll eine Bewertung der Leistung und Abschätzung zu den Kosten im Rahmen der verschiedenen Computing-Modelle mit und ohne Data-Lake-Szenario erstellt werden.
Die bei den verschiedenen Tests gewonnenen Erfahrungen, sowie auftretende Probleme, müssen im Detail ausgewertet werden und gegebenenfalls in weitere Optimierungen und Entwicklungen der betroffenen Dienste einfließen, d.h. in die Arbeiten im Themengebiet I und II zurückgespiegelt werden.

Arbeitspaket 2: Spezifische Anpassung der Dienste an komplexe Workflows und Nutzung spezieller Technologien für die Analyse wissenschaftlicher Daten Ein weiterer zentraler Aspekt sind Tests und Anpassungen der entwickelten Systeme an die unterschiedlichen Anwendungen in den verschiedenen Wissenschaftsbereichen bzw. Projekten. Die konkreten Anforderungen an Ressourcen (Speicher, Rechenzeit, seriell oder parallel, IO-Last, Netzwerkanbindung, GPU-Nutzung) unterscheiden sich stark, je nachdem ob es um Simulation, Rekonstruktion oder Analyse geht. Zusätzlich haben die verschiedenen Projekte unterschiedliche Anforderungen an die Hardware (z.B. GPU für die Simulation oder Rekonstruktion der Daten). Mit diesen Unterschieden in den Arbeitsabläufen, in Bezug auf die Hardwareanforderungen, muss das Workflowmanagement umgehen können. Das Arbeitspaket soll Lösungen für die Anpassung und Integration dieser Anforderungen erarbeiten. Dies sind zum Teil allgemeine Lösungen, aber experimentspezifische Anpassungen sind ebenfalls erforderlich. Insgesamt muss eine effiziente Nutzung der Ressourcen für die entsprechenden Anwendungen gefunden werden.
In diesem Zusammenhang sollen auch spezielle Workflows, wie die schnelle Analyse mit sehr großen Datenmengen, behandelt werden. Ziel ist die Reduktion der Zeit für einen typischen Arbeitszyklus zum Prozessieren von 1 TB Daten von 1 Tag auf 1 Stunde. Dies erfordert die Kombination von a) neuer Software für Parallelisierung und Caching, b) rigorose Vektorisierung der Algorithmen anstelle der Verwendung von Event-Loops und c) Hardwareausstattung mit SSD für einen extrem hohen Datendurchsatz.
Ein weiteres Ziel des Arbeitspakets ist die Integration der National Analysis Facility am DESY für die breite und damit sehr heterogene Nutzung für die Analyse verschiedenster Experimente und Nutzerkreise, sowie die Nutzung der Dienste bei der National Analysis Facility bei GSI und dem sich im Aufbau befindlichen Daten- und Analysezentrum der Astroteilchenphysik (KIT und DESY in Zeuthen).

Arbeitspaket 3: Support Für eine erfolgreiche Umsetzung und Erprobung der Werkzeuge auch an Standorten, die keine dedizierte Expertise vor Ort haben, ist eine aktive Unterstützung bei Installation, Betrieb und Wartung der zentralen Werkzeuge notwendig. Dazu müssen ausführliche Dokumentationen oder auch Schulungsunterlagen erstellt werden und kompetente Ansprechpartner bereit stehen. Als weitere Maßnahme sind Workshops mit interessierten Nutzern geplant, in denen Installation, Betrieb und Überwachungswerkzeuge in einer Hands-on-Umgebung demonstriert werden. Schulungen auf etablierten Computing-Schulen oder aus Anlass der Jahrestreffen der großen Kollaborationen sollen vorbereitet und durchgeführt werden, um die Verbreitung der entwickelten Werkzeuge zu fördern. Zur Sicherstellung des stabilen und dauerhaften Betriebs wird die Einrichtung eines standortübergreifenden “Support Teams” angestrebt. Das Team soll aus mindestens zwei Experten für jede Komponente bestehen. Klar definierte Ansprechpartner bei Fragen und Problemen bei Installation und Betrieb tragen erheblich zur Effizienz des Gesamtprojektes bei.

3.3.2 Beteiligte Institute, Koordination, beantragte Mittel

Die beteiligten Partner im Themenbereich, sowie die beantragten Mittel sind in Tabelle 3 aufgeführt. Die Koordination des Themenbereichs wird übernommen von Prof. Christian Zeitnitz und Dr. Günter Duckeck.

Expertise: Alle beteiligten Partner haben direkt oder indirekt Erfahrung beim Betrieb von größeren Rechenclustern. Die Standorte Aachen, CERN, DESY, GSI, KIT, Wuppertal, München, Freiburg und Göttingen betreiben ein WLCG-Tier-Zentrum. Die Standorte Mainz und Hamburg betreiben große Tier-3 Installationen. Frankfurt verfügt über Erfahrung beim Betrieb hybrider CPU/GPU-Cluster.
Diese Standorte kennen sich daher auch mit der Installation und dem Betrieb der benötigten Dienste aus. Sie sind daher für die Durchführung der oben beschriebenen Tests prädestiniert. Die Betreiber der Tier-Zentren besitzen auch die Expertise für die verschiedenen Workflows der Experimente.
Im Rahmen der Pilotmaßnahme ErUM-Data haben die Standorte Aachen, KIT, Frankfurt, GSI, Freiburg und München bereits viel Erfahrung bei der Installation der entwickelten Dienste für den Tier-2 Betrieb, Daten-Caching und der Nutzung von opportunistischen Ressourcen gewonnen. Container-Technologien sind die Basis für viele der oben beschriebenen Dienste. Hier bestehen langjährige Erfahrungen in Göttingen, Frankfurt, GSI, Hamburg, KIT und Wuppertal.

4 Vernetzung der Themenbereiche

Aus der Konzipierung dieses Antrages ist ersichtlich, dass die Themenbereiche eng vernetzt sind. Die in Themenbereich I entwickelten Werkzeuge zur Erschließung heterogener Ressourcen werden erst durch die effiziente Anbindung der Massenspeicher (”Data-Lakes”), die in Themenbereich II entwickelt werden, für den Produktionsbetrieb nutzbar. Die Anpassung

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

Tabelle 3: Standorte, beantragte Mittel und assoziierte Partner im Themenbereich III. Eine Tabelle mit der Gesamtübersicht aller Themenbereiche findet sich am Ende des Dokuments.

und Integration, sowie die kombinierten Tests in Themenbereich III müssen in enger Abstimmung mit den Entwicklern erfolgen. Viele beteiligte Arbeitsgruppen sind in mehr als einem Themenbereich aktiv, so dass die Vernetzung offensichtlich ist.

Die enge Zusammenarbeit zwischen den beteiligten Partnern, sowie auch weiteren deutschen und internationalen Gruppen zur erfolgreichen Durchführung des hier beantragten Projekts ist zwingend notwendig. Geplant sind dazu regelmäßige Treffen und Workshops der Arbeitspakete und Themenbereiche. Diese Treffen können an den Projektstandorten, am CERN und online stattfinden. Darüber hinaus sind Berichte über den Projektfortschritt auf internationalen Konferenzen geplant. Einmal pro Jahr soll ein Workshop mit allen beteiligten Partnern und potentiellen Nutzern der entwickelten Werkzeuge abgehalten werden. Hierdurch wird sichergestellt, dass sich die geplanten Arbeiten durch engen Kontakt zu den Nutzern in die richtige Richtung entwickeln. Auch die Verbreitung der neuen Werkzeuge soll hierdurch gefördert werden.

Die interdisziplinäre Vernetzung wird durch diese Projekte außerdem innerhalb der einzelnen Standorte gefördert. Dies wird unter anderem am Beispiel KIT deutlich, wo Teilchenphysik, Astroteilchenphysik und die Informatik in gemeinsame Projekte eingebunden sind.

Um die Vernetzung über den gesamten Verbund zu fördern wird vom Verbundsprecher (A. Schmidt, RWTH Aachen) für diesen Zweck eine halbe Stelle beantragt.

5 Bezahlung von Doktorandinnen und Doktoranden in BMBF–Projekten

Das Komitee für Elementarteilchenphysik (KET) und das Komitee für Hadronen und Kernphysik (KHuK) empfiehlt in einem Schreiben vom Mai 2020 die Bezahlung von Doktorand*innen in BMBF-geförderten Projekten von in der Regel 50% einer Wissenschaftler*innen Stelle (E13) auf 67% in der kommenden Förderperiode anzuheben. Es wird dabei betont, dass eine höhere Bezahlung der Doktorand*innen in BMBF-Projekten nicht zu einer Reduzierung des insgesamt BMBF-finanzierten wissenschaftlichen Personals in diesen Projekten führen darf. KET und KHuK bitten daher das BMBF, speziell für diesen Zweck zusätzliche Mittel bereitzustellen. Der Wortlaut der Empfehlung ist unter `https://www.ketweb.de/e199639/e308231/Text-PhD-v2.3.pdf` einsehbar.

Alle an den ALICE, ATLAS, BELLE II, CMS und LHCb Verbundanträgen beteiligten Arbeitsgruppen schließen sich ausdrücklich dieser Empfehlung an und beantragen deshalb 67% einer Wissenschaftler*innen Stelle für die Bezahlung von Doktorand*innen.

6 Zusammenfassung

Im beantragten Projekt sollen die Weichen gestellt werden, um den zukünftigen Herausforderungen im wissenschaftlichen Computing in der Teilchen-, Astroteilchen-, sowie Hadronen- und Kernphysik zu begegnen. Damit soll die Exzellenz der deutschen Grundlagenforschung weiterhin gesichert werden. Im Projekt sollen Werkzeuge und Technologien zur Einbindung und Nutzbarmachung heterogener Computing-Ressourcen sowie der dazugehörigen Massenspeicher - Infrastrukturen entwickelt werden. Darüber hinaus sollen die neuen Technologien an konkrete Zielsysteme angepasst und in kombinierten Tests systematisch erprobt werden. Eine Übersicht der beteiligten Partner und der beantragten Mittel ist in Tabelle 4 gegeben.

Literatur

[1] *Zusammenfassung der Strategiediskussion zum Computing in der HL-LHC-Ära*. `https://indico.physik.uni-muenchen.de/event/33/attachments/142/242/Abschlussdokument.pdf`. Version: Mai 2020

[2] Bird, I: Status of the WLCG Project, including Financial Status / CERN. Version: Aug 2016. `https://cds.cern.ch/record/2210410`. Geneva, Aug 2016 (CERN-RRB-2016-124). – Forschungsbericht

[3] *CernVM*. `https://cernvm.cern.ch`. Version: 2017, Abruf: 2017-09-22

[4] *The XRootD software framework*. `http://xrootd.org`. Version: 2017, Abruf: 2017-09-22

[5] Fischer, Max ; Kuehn, Eileen ; Giffels, Manuel ; Schnepf, Matthias ; Kroboth, Stefan ; Freyermuth, Oliver: COBalD - The opportunistic balancing daemon. (2020), Apr. http://dx.doi.org/10.5281/zenodo.1887872. – DOI 10.5281/zenodo.1887872

[6] Giffels, Manuel ; Schnepf, Matthias ; Kuehn, Eileen ; Kroboth, Stefan ; Caspart, Rene ; Cube, Ralf F. ; Fischer, Max ; Wienemann, Peter: TARDIS - Transparent Adaptive Resource Dynamic Integration System. (2020), Jun. http://dx.doi.org/10.5281/zenodo.2240605. – DOI 10.5281/zenodo.2240605

[7] Knedlik, Jan ; Kramp, Paul ; Schwarz, Kilian ; Kollegger, Thorsten: XRootD plug-in based solutions for site specific requirements. (2019), Sep. http://dx.doi.org/10.1051/epjconf/201921404005. – DOI 10.1051/epjconf/201921404005

Standort	PI	FTE	Experimente	Themenbereich I	II	III	Anteilige Computing-Ressourcen verfügbar für Tests
RWTH Aachen	A. Schmidt / M. Erdmann / A. Stahl	2	CMS Einstein Teleskop			X	WLCG-Tier-2, HPC der RWTH
KIT	G. Quast / A. Streit / R. Engel	3	CMS, Auger	X	X	X	WLCG-Tier-3, HPC des KIT
U Bonn	P. Bechtle	1	ATLAS/Belle II	X			WLCG-Tier-3, HPC der Uni Bonn
GU Frankfurt	V. Lindenstruth	2	ALICE/CBM	X	X	X	Goethe HLR
LMU München	T. Kuhr / G. Duckeck	2	ATLAS/Belle II		X	X	WLCG-Tier-2/3, HPC (SuperMUC)
U Freiburg	M. Schumacher	2	ATLAS	X		X	WLCG-Tier2/3, HPC (NEMO)
U Göttingen	A. Quadt	1.66	ATLAS	X	X	X	WLCG-Tier-2/3, HPC (HLRN, SCC)
U Mainz	F. Maas / A. Brinkmann	1.66	PANDA		X	X	
U Wuppertal	C. Zeitnitz	1.66	ATLAS	X		X	WLCG-Tier-2/3
U Hamburg	J. Haller	1.33	CMS		X	X	Tier-3, HPC der U Hamburg
Assoziiert							
GSI	K. Schwarz	-	ALICE	X	X	X	Green IT Cube
DESY	V. Gülzow	-	verschiedene	X	X	X	
CERN	M. Elsing	-	ATLAS		X		
GridKa	A. Petzold	-	verschiedene	X	X	X	GridKa WLCG-Tier-1

Tabelle 4: Standorte, beantragte Mittel und assoziierte Partner in allen Themenbereichen.