Entwurf einer Infrastruktur für den Datenaustausch großer Forschungsdatenmengen mittels WebDAV, FTS3 und OIDC

Martin Baumann[1], Frauke Bösert[2], Sven Siebler[1], Paul Skopnik[2] und Jan Erik Sundermann[2]

[1]Universitätsrechenzentrum, Universität Heidelberg
[2]Steinbuch Centre for Computing, Karlsruher Institut für Technologie

Standortübergreifende Zusammenarbeit innerhalb wissenschaftlicher Communities im Umfeld föderierter Speicher- und Compute-Umgebungen erfordert häufig den Transfer großer Datenmengen zwischen Zentren und Speichersystemen. Vorgestellt wird eine prototypische Implementierung einer Infrastruktur, die auf einem mit Plugins erweiterten Apache-Webserver für den Datenaustausch mittels WebDAV basiert. Die Server ermöglichen dabei den Zugriff auf Speichersysteme und ergänzen diese um ein weiteres Zugriffsprotokoll, welches die Anbindung an FTS3 unterstützt.

Die erforderlichen Erweiterungen und erste Ergebnisse zum Transfer von Daten mittels WebDAV werden vorgestellt.

1 Einleitung

Die Zusammenarbeit in wissenschaftlichen Communities erfordert oft Transfers großer Datenmengen zwischen Zentren und Speichersystemen sowie authentifizierte Datenzugriffe. Die einfache Einbindung existierender Speichersysteme von Hochleistungsrechnern, Archivierungsdiensten oder Repositorien ist wünschenswert wenn nicht sogar notwendig, da auf diesen Systemen die großen Datenmengen erzeugt, analysiert oder später archiviert bzw. veröffentlicht werden. Der im folgenden vorgestellte Ansatz unterstützt bei der Lösung dieser Anforderungen und greift existierende bzw. sich aktuell in intensiver Erprobung befindliche Konzepte zum automatisierten und effizienten Datentransfer auf Basis von FTS3 und Standard-Protokollen [1, 2] aus dem Umfeld des LHC-Computings auf und ergänzt diese um technische Lösungen, die es ermöglichen sollen, bereits existierende und heterogene Speichersysteme in die föderierte Infrastruktur zu integrieren.

Die Durchführung entsprechender Transfers setzt ein geeignetes Netzwerkprotokoll, eine flexibel nutzbare Verwaltungsschicht für die Steuerung und Überwachung der Übertragungen und eine passende Authentifizierung voraus. Für den Transfer großer Datenmengen

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13753 (CC BY-SA 4.0)

hat sich das HTTP-basierte Netwerkprotokoll WebDAV als vielversprechende Technologie in verschiedenen Anwendungsfeldern bereits bewährt. Zur Steuerung und Überwachung von Transfers großer Datenmengen kann die Open-Source Software FTS3 genannt werden, die zur Übertragung der Daten des Teilchenbeschleunigers am CERN seit längerem produktiv eingesetzt wird. FTS3 fungiert dabei als zentrale koordinierende Stelle oder "dritte Partei", die Datentransfers zwischen zwei Speichersystemen initiieren kann. Der Datenfluss erfolgt dabei immer direkt zwischen den beiden beteiligten Speichersystemen. Mit OpenID Connect ist eine auf OAUTH2 aufbauende Authentifizierungsschicht gegeben, auf Basis derer eine token-basierte Authentifizierung ermöglicht wird. Mittels dieser Authentifizierungstokens können Berechtigungen für Datentransfers vorübergehend, beispielsweise an den FTS3-Transferdienst, delegiert werden.

2 Involvierte Technologien

FTS3: Die Open-Source Software FTS3 (`https://fts.web.cern.ch/fts/`) [3] wurde entwickelt, um Datenmengen im Petabyte-Bereich vom Teilchenbeschleuniger Large Hadron Collider am Europäischen Kernforschungszentrum CERN global zu verteilen und ist seit längerem produktiv im Einsatz. Sie ermöglicht es, Datenübertragungen effizient zu planen und die vorhandenen Netzwerkressourcen optimal einzusetzen. Zur Steuerung bietet FTS3 ein Kommandozeileninterface (CLI), eine REST API und mit WebFTS (`https://webfts.cern.ch`) ein web-basiertes Nutzerinterface. Die Authentifizierung erfolgt mittels x.509 Zertifikaten oder per OpenID Connect (OIDC/OAUTH2). Als Datenübertragungsprotokolle werden HTTPS/WebDAV, XrootD und GridFTP unterstützt. Mit FTS3 kann ein Dienst zur Orchestrierung von Third-Party-Datentransfers (TPC) zwischen verschiedenen Speichersystemen (URIs) betrieben werden.

WebDAV: Das im RFC 4918 spezifizierte Netzwerkprotokoll WebDAV (`http://webdav.org/`) basiert auf HTTP/HTTPS und erweitert dieses Protokoll um einen Satz neuer Befehle. Diese Befehle erlauben es beispielsweise, Eigenschaften von Dateien, Verzeichnissen oder Verzeichnisstrukturen abzurufen, Verzeichnisse zu erstellen, sowie Dateien oder Verzeichnisse zu kopieren, zu verschieben oder zu löschen. Als HTTP-basiertes Protokoll ist es standardisiert und darüber hinaus weit verbreitet. Entsprechende Implementierungen stehen für die üblichen Betriebssysteme (Win/Lin/Mac auch Android/iOS) wie auch in den gängigen Webservern zur Verfügung. Das WebDAV-Protokoll ermöglicht die Übertragung von einzelnen Dateien wie auch von ganzen Verzeichnissen. Die Verwendung von WebDAV ist gerade für standortübergreifende Verbindungen interessant, da der benötigte Port 443 als Standardport für HTTPS-Verbindungen in der Regel bereits freigegeben ist und somit meist keine speziellen Firewallanpassungen erforderlich werden. WebDAV-Verbindungen lassen sich abhängig von der gewählten Serverimplementierung auf verschiedene Weisen authentifizieren und es ist im Allgemeinen möglich, eine token-basierte Authentifizierung zu realisieren, wie sie bei dem hier beschriebenen Ansatz zur Integration mit FTS3 favorisiert wird.

OpenID Connect (OIDC): OIDC (`https://openid.net/connect/`) ist eine auf OAuth 2.0 basierte, inzwischen sehr weit verbreitete Authentifizierungsschicht. OIDC ermöglicht die Authentifizierung von Nutzern bei Diensten, ohne dass diese Passwörter austauschen oder verwalten müssen. Durch den Austausch von Authentifizierungstokens können Berechtigungen für Datentransfers vorübergehend beispielsweise an den FTS3-Transferdienst delegiert werden. Bei Verwendung geeigneter Werkzeuge [4, 5] können diese ähnlich wie ssh-Keys in übliche Workflows beim Datenzugriff integriert werden.

3 Aufbau des Prototyps

Es wurde begonnen, einen Prototyp der geplanten Infrastruktur aufzubauen. Mit diesem Prototyp sollte zunächst die Durchführbarkeit von TPC und die Funktionalität im Zusammenspiel der involvierten Technologien demonstriert werden und er sollte anschließend weiteren Untersuchungen dienen. Der Prototyp besteht aus den bereits genannten Technologien WebDAV, FTS3 und OIDC, die geeignet erweitert bzw. modifiziert wurden, sowie einzelnen Speicherendpunkten.

Zur Bereitstellung der Protokolle für den Datenzugriff wurden WebDAV-Server auf Basis des Apache-HTTP-Servers mit erweiterten und neuen Modulen verwendet. Nutzerdaten lagen in lokalen Dateisystemen vor. Der Ansatz lässt sich auf beliebige POSIX-Dateisysteme generalisieren, wie sie beispielsweise in den zentralen Speichersystemen an HPC-Clustern eingesetzt werden. Um den Apache WebDAV-Server auf beliebigen POSIX-Dateisystemen einsetzen zu können, ohne die existierenden Eigentümer und Berechtigungsstrukturen ändern zu müssen, wurde dieser so konfiguriert, dass WebDAV-Befehle mit der UID des authentifizierten Nutzers ausgeführt werden. Dazu wird das Apache-Modul mpm-itk [6] eingesetzt, welches es ermöglicht jede HTTP/HTTPS-Anfrage einzeln unter möglicherweise verschiedenen POSIX-Benutzern zu verarbeiten. Zu diesem Zweck verändert mpm-itk den Anfragen-Verarbeitungs-Mechanismus von Apache: Statt Anfragen von langlebigen Prozessen mit einer statisch konfigurierten UID verarbeiten zu lassen, wird jede Anfrage in einem eigenen Prozess verarbeitet, wobei dessen UID anfragespezifisch abgeleitet wird. Der zusätzliche Aufwand, einen neuen Prozess pro Anfrage zu starten, ist unbedeutend im Vergleich zum Datentransfer von großen Dateien. Für den WebDAV-Anwendungsfall wurde das Modul weiter angepasst, so dass die Entscheidung über die UID erst nach erfolgreicher Authentifizierung des Nutzers passiert. Mit diesem Setup ist es möglich, authentifizierten Nutzern Zugriff auf existierende POSIX-Speichersysteme zu geben und dabei alle durch das Dateisystem unterstützten Berechtigungsstrukturen inkl. erweiterter ACLs zu berücksichtigen.

Die Nutzerauthentifizierung erfolgt mittels OAUTH2 Bearer-Tokens und OpenID Connect [7, 8]. Die so authentifizierten Nutzer werden mit Hilfe eines dedizierten Moduls lokalen POSIX-Benutzern zugeordnet. Unter deren UID wird die Anfrage dann vearbeitet. Die eingesetzten Speichersysteme verwenden LDAP-Facaden und basieren damit vollständig auf bwIDM, dem föderiertes Identitätsmanagement in Baden-Württemberg [7].

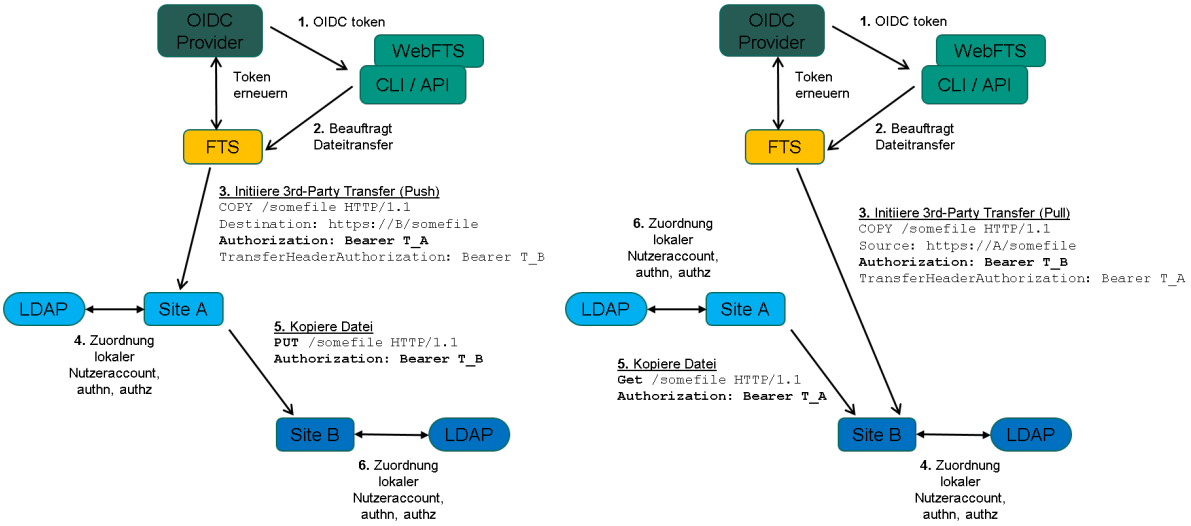

Abbildung 1: Third-Party-Dateitransfers mit WebDAV / FTS3 / OIDC im Push- (links) und Pull-Modus (rechts).

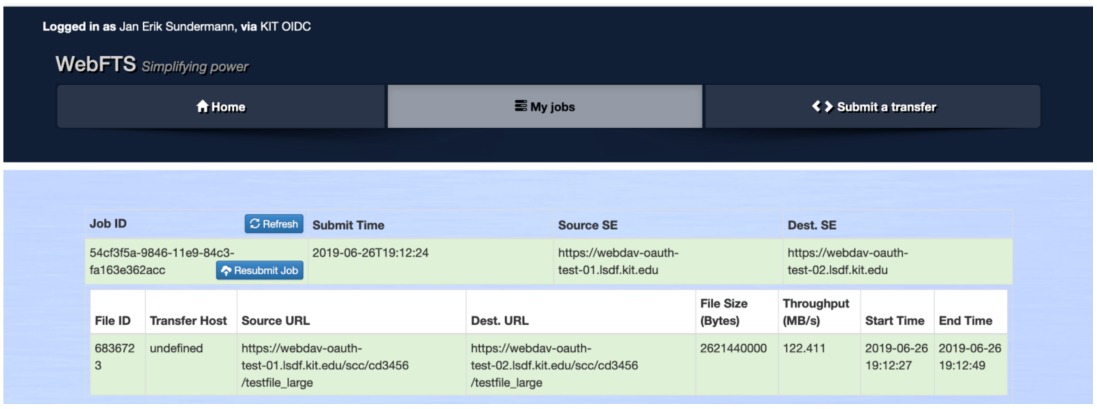

Abbildung 2: WebFTS-Instanz mit Beispiel eines erfolgreichen Transfers zwischen zwei Endpunkten.

Für die Integration eines vorhandenen Speichersystems an FTS3 muss der entsprechende Speicherendpunkt einen erweiterten Befehlssatz von WebDAV verstehen, der es ihm ermöglicht, TPC durchzuführen (siehe [10]). Zur Integration in den beschriebenen Prototypen wurde das existierende Apache WebDAV-Modul um den TPC-Befehl COPY sowie um Performance-Marker erweitert. Die Implementierung verwendet die GFAL2-Bibliothek [11]. Für diese Evaluation wurde ein existierender FTS3-Server am CERN verwendet, sowie eine neue WebFTS-Instanz am KIT aufgesetzt. Abbildung 1 illustriert den Ablauf eines von FTS initiierten TPC im Push- und im Pull-Modus. Nur einer der involvierten Speicherendpunkte muss einen WebDAV-Zugang bereit stellen, der den erweiterten Befehlssatz für TPC unterstützt.

4 Ergebnisse

Auf Basis des zuvor beschriebenen Prototypen wurde zunächst die grundsätzliche Funktionalität erprobt. Hierbei wurden mit Hilfe von WebFTS standortübergreifende TPC durchgeführt (siehe Abb. 2). Weiterhin wurden erste vergleichende Performancemessungen des WebDAV- und SFTP-Protokolls zwischen Heidelberg und Karlsruhe durchgeführt. Hierfür wurde das Tool rclone (`https://rclone.org/`) mit einem synthetischen Datensatz verwendet, bestehend aus je 147 GB Daten verschiedener Dateigröße (16 MB - 10 GB; 9362 - 14 Dateien). Die Benchmarkergebnisse für die Transfers mit WebDAV und SFTP sind in Abb. 3 dargestellt.

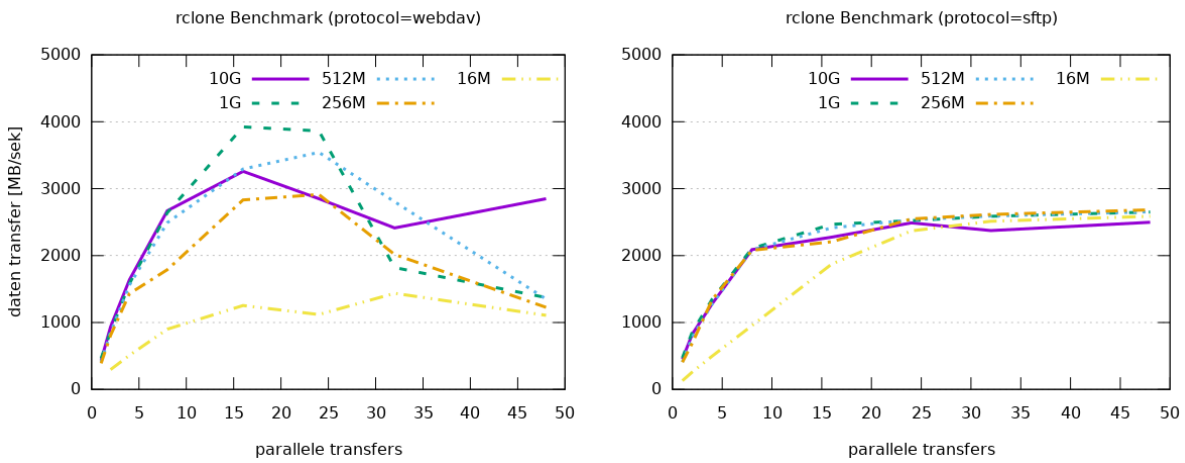

Abbildung 3: rclone Benchmark auf einem synthetischen Datensatz (je 147 GB einer Dateigröße).

In beiden Fällen kann ein deutlicher Performancegewinn durch Verwendung paralleler Transfers erreicht werden. Bei WebDAV fällt dieser Effekt zum einen stärker aus, zum anderen wird (bei bis zu 25 Transfers) mit 3-4 GB/s auch eine höhere Maximalgeschwindigkeit erreicht. Der Performanceeinbruch ab ca. 25 Transfers ist vermutlich auf serverseitige Parameter zurückzuführen, so dass durch weitere Optimierung auch eine höhere Skalierung zu erwarten ist. Bei Verwendung des SFTP-Protokolls erkennt man im Vergleich, dass das mögliche Maximum von 2.2-2.5 GB/s bereits durch 8-10 Transfers erreicht wird. Lediglich bei kleineren Dateigrößen skaliert SFTP deutlich besser als WebDAV, was wahrscheinlich auf den Verbindungsoverhead durch die unterschiedliche Dateianzahl zurückzuführen ist.

WebDAV ist für den skalierenden Transfer von großen Datenmengen, auch standortübergreifend, geeignet. Insbesondere die Möglichkeit des automatisierten Datentransfers durch TPC bietet hierbei einen deutlichen Mehrwert für die Zusammenarbeit beim föderierten Datenaustausch.

5 Ausblick

Für den vorbereitenden Einsatz des Prototypen in Produktivumgebungen und die Verbesserung der Performanceskalierung sind im Weiteren folgende Aktivitäten vorgesehen: Optimierung verwendeter Apache-Webserver-Parameter, Verbesserung der TPC-Implementierung in WebDAV, Anbindung weiterer Speicherendpunkte und OIDC-Provider. Zusätzlich ist die Erprobung des WebDAV-Protokolls mit gängigen Anwendungen und Nutzerworkflows geplant, z.B. für die Datenanalyse mit Scikit-Learn und Jupyter Notebooks. Darüber hinaus soll evaluiert werden, wie sich die so aufgesetzte Infrastruktur zum automatisierten und regelbasierten Datentransfer z.B. mit Datenmanagementwerkzeugen wie Rucio [12] einsetzen lässt.

Danksagung

Die vorgestellten Ergebnisse wurden im Rahmen des Projekts bwHPC-S5 erarbeitet, das durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK) gefördert wird.

Die Benchmarkmessungen wurden auf den Speicherdiensten “LSDF Online Storage” (Karlsruhe) und “SDS@hd” (Heidelberg) durchgeführt, die vom MWK und der Deutschen Forschungsgemeinschaft (DFG) gefördert sind (INST 35/1314-1 FUGG, INST 35/1503-1 FUGG).

Literaturverzeichnis

[1] B. Bockelman, A. Ceccanti, F. Furano1, P. Millar, D. Litvintsev and A. Forti. “Third-party transfers in WLCG using HTTP”, EPJ Web Conf., Volume 245 (2020), 24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019)

[2] B. Bockelman, A. Hanushevsky, O. Keeble, M. Lassnig, P. Millar, D. Weitzel, W. Yang. “Bootstrapping a New LHC Data Transfer Ecosystem”, EPJ Web Conf., Volume 214 (2019), 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018)

[3] Kiryanov, A., T. A. Ayllon, and O. Keeble. “FTS3 / WebFTS – A Powerful File Transfer Service for Scientific Communities”, Procedia Computer Science, volume 66 (2015): 670-678.

[4] OIDC-Agent Projektseite, `https://indigo-dc.gitbook.io/oidc-agent/` [abgerufen 27. Juli 2021].

[5] G. Zachmann. “OIDC-Agent: Managing OpenID Connect Tokens on the Command Line”, In: Becker, M. (Hrsg.), SKILL 2018 - Studierendenkonferenz Informatik. Bonn: Gesellschaft für Informatik e.V. (S. 11-21).

[6] The Apache 2 ITK MPM, http://mpm-itk.sesse.net/,[abgerufen 27. Juli 2021].

[7] Apache-Modul mod_oauth2, https://github.com/zmartzone/mod_oauth2 [abgerufen 27. Juli 2021].

[8] Apache-Modul mod_auth_openidc, https://github.com/zmartzone/mod_auth_openidc,[abgerufen 27. Juli 2021].

[9] J. Köhler, S. Labitzke, M. Simon, T. Dussa, M. Nussbaumer, H. Hartenstein. “bwIDM – Federated Access to IT-Based Services at the Universities of the State of Baden-Württemberg”, De Gruyter, Online erschienen: 25. Januar 2014, https://doi.org/10.1515/pik-2013-0025.

[10] HTTP/WebDAV Third-Party-Copy Technical Details, CERN Wiki(25. März 2020): https://twiki.cern.ch/twiki/bin/view/LCG/HttpTpcTechnical [abgerufen 27. Juli 2021].

[11] Grid File Access Library (GFAL2), https://dmc-docs.web.cern.ch/dmc-docs/gfal2/gfal2.html,[abgerufen 27. Juli 2021].

[12] M. Barisits, T. Beermann, F. Berghaus et al. “Rucio: Scientific Data Management”, Computing and Software for Big Science volume 3, Article number: 11 (2019), https://doi.org/10.1007/s41781-019-0026-3.