Ausarbeitung

Scientific Cloud Computing

vorgelegt von

Moritz Hartkopf

Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Arbeitsbereich Wissenschaftliches Rechnen
Proseminar „Speicher- und Dateisysteme”

Studiengang: Wirtschaftsinformatik
Matrikelnummer: 7052417
Betreuer: Dr. Hermann Lenhart

Hamburg, 2019-06-03

Inhaltsverzeichnis

1 Einleitung

Cloud Technologie ist eines der größten Themen der letzten Jahre in der IT. Sie bringt innovative Möglichkeiten für die Wirtschaft mit sich und lässt immer mehr neue Geschäftsmodelle entstehen. Doch nicht nur für kommerzielle Zwecke bieten Clouds ein riesiges Potenzial. Auch in der Wissenschaft spielt Cloud Technologie eine immer wichtigere Rolle beim Umgang mit den wachsenden Datenmengen im Rahmen der Digitalisierung und des technischen Fortschritts in der Forschung.

Im Folgenden erläutere ich inwiefern Cloud Technologie das wissenschaftliche Arbeiten beeinflussen kann und welche Potenziale sie birgt. Hierzu führe ich einige Beispiele an, wie Clouds bereits in der Forschung zum Einsatz kommen, gebe einen Einblick über grundsätzliche Technologien im Hintergrund und entwickele Anforderungen an Scientific Clouds. Außerdem erläutere ich deren Nutzen und zeige Risiken auf, welche der Einsatz von Cloud Technologie mit sich bringen kann.

2 Beispiele

Im folgenden werden einige Beispiele vorgestellt, die veranschaulichen welche Anforderungen Scientific Clouds erfüllen müssen. Dieses Anforderungen lassen sich in drei Hauptkategorien gliedern: Den Umgang mit großen Mengen an wissenschaftlichen Daten, das Bilden gemeinsamer Plattformen für wissenschaftliche Publikationen und eine ähnliche Verwaltung wie in gewöhnlichen Clouds für die Öffentlichkeit, jedoch mit erhöhten Sicherheitsanforderungen.

2.1 Max Planck Gesellschaft

Die Max Planck Gesellschaft ist eine deutsche Institutionen im Bereich der Grundlagenforschung und macht sich Science Clouds für wissenschaftliche Zwecke zu Nutze. Beispielweise der GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung) Cloud Share ist ein Dienst für Datenaustausch und -synchronisation. Er ermöglicht, dass Dateien zwischen unterschiedlichen Geräten wie dem Computer im Büro und dem eigenen Laptop synchronisiert werden können. Der Dienst stellt außerdem eine alternative Möglichkeit zum Austausch von großen Email Anhängen dar und die Zusammenarbeit mit anderen Forschern an einer oder mehreren Dateien wird durch schnellen und ortsunabhängigen Zugriff vereinfacht. Man kann Daten offline bearbeiten, die aktualisiert werden, sobald man wieder online ist. Auch virtuelle Server unterstützen die Forschungsarbeiten. Durch diese kann je nach Bedarf beliebig viel Rechenleistung binnen Sekunden in Anspruch genommen werden.

2.2 DESY

Das DESY ist ein Forschungszentrum für naturwissenschaftliche Grundlagenforschung mit Sitz in Hamburg und Zeuthen. Auch dort wird Cloud Computing zur Unterstützung der Forscher genutzt. Die Daten werden in der DESY Infrastruktur gespeichert, weshalb auch vertrauliche Daten bedenkenlos hochgeladen werden können. Hauptziel der DESY Cloud ist es, eine alternative Plattform, zu Google Drive und ähnlichen Diensten zum speichern von Daten bereitzustellen. Außerdem soll sie das Synchronisieren von Dateien auf verschiedenen Geräten ermöglichen und zum einfachen Austausch der Daten unter Kollegen, die gemeinsam an Projekten arbeiten, beitragen [24]. Das DESY-Rechenzentrum fasste 2015 insgesamt rund elf Petabyte an wissenschaftlichen Daten [26]. Gespeichert werden diese auf mehreren tausend Festplatten und gesichert auf Magnetbändern. Bei Experimenten mit dem DESY Teilchenbeschleuniger fallen sehr große Datenmengen an.

Bereits 2015 erzeugte der DESY Teilchenbeschleuniger PETRA III ca. 10 Terabyte an wissenschaftlichen Daten pro Sekunde [25]. Aktuellen Angaben des DESY zu Folge, soll der Röntgenlaser XFEL vom DESY momentan 3000 und ab 2020 sogar bis zu 27 000 Röntgenlaserblitze pro Sekunde produzieren [27]. Dabei entstehen ebenfalls enorme Datenmengen. Bereits in den ersten fünf Monaten Forschung sind rund 700 Terabyte an Daten mit der Superkamera generiert worden [16]. Um die Datenfluten zu bewältigen, arbeitet das DESY bei den Datenanalysen mit dem CERN als Kooperationspartner zusammen.

2.3 CERN

Das CERN ist ein weiteres gutes Beispiel für die Nutzung von Cloud Technologie in der Wissenschaft. CERN ist die Europäische Organisation für Kernforschung und hat ihren Sitz in Genf in der Schweiz. Es ist eines der am meisten Rechenleistung benötigenden Forschungszentren in der Wissenschaft.

Bisher war vor allem das LHC Computing Grid für die Verarbeitung und Analyse zuständig, doch mittlerweile wird auch vermehrt auf wissenschaftliche Clouds gesetzt um die stetig wachsende Datenflut zu bewältigen. Clouds ermöglichen eine deutlich einfachere Benutzung als das Computing Grid und dienen dazu auch weniger technisch befähigte Teilchenforscher anzusprechen wie zum Beispiel Chemiker. Bei Experimenten mit dem Large Hadron Collider, dem weltgrößten Teilchenbeschleuniger, werden pro Sekunde bis zu eine Millionen Teilchenkollisionen durchgeführt. Dies entspricht etwa einem Petabyte an Daten, welche die Sensoren liefern. Zum Vergleich generiert der Petra III Teilchenbeschleuniger beim DESY nur etwa 10 Terabyte pro Sekunde [25]. LHC Experimente produzieren jährlich eine Datenmenge von ca. 25 Petabyte, trotz späterer Filterung [25]. Diese Menge entspricht mehr als einer Millionen DVDs pro Jahr [31]. In allen LHC-Experimenten zusammen liefern rund 150 Millionen Sensoren vierzig Millionen Mal pro Sekunde Daten [29].

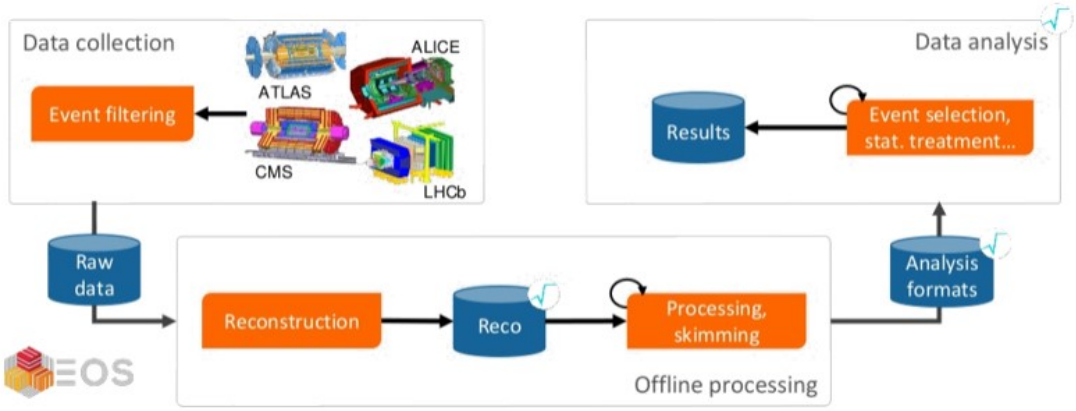
Das Speichersystem EOS vom CERN fasste 2015 ca. 55 Petabyte während im Rechenzentrum vom DESY zeitgleich etwa 11 Petabyte an wissenschaftlichen Daten lagerten [25]. Mittlerweile liegen dort mehr als 2 Millionen Dateien und die Gesamtspeicherkapazität überschreitet 250 Petabyte. Für die Langzeitdatenspeicherung werden im CERN momentan noch vor allem Magnetbänder verwendet. Der Archivinhalt wird langsam auf leistungsfähigere Higher Density Tapes.

Seit dem 15.11.2018 lässt sich das CERN auch durch Commercial Clouds unterstützen. Zu den CERN OpenLab Kollab Partnern gehört seit dem auch Google mit der Google Cloud. Google verspricht, dass immer auf Daten zugegriffen werden kann, auch wenn mal ein Rechenzentrum ausfallen sollte gibt es eine Kopie auf der Cloud. Mit dem High-Luminosity LHC, welcher ab 2026 Einsatzbereit sein soll, rechnen die Mitarbeiter von Google damit, dass ein Exabyte an Speicher nötig sein wird, um die entstehenden

Daten zu sichern.

2.3.1 CERN Datenpipeline

Im ersten Schritt der CERN Daten Pipeline werden Daten gesammelt. Ein Großteil der Daten von den Experimenten wird herausgefiltert. Bei Experimenten mit dem ATLAS Teilchendetektor bleiben nach dem Aussortieren "nur" noch ein Gigabyte an Daten pro Sekunde von den ursprünglichen ein Petabyte pro Sekunde übrig, welche insgesamt bei den Experimenten anfallen [26]. Anschließend werden die Rohdaten rekonstruiert. Dabei wird aus den einzelnen Messwerten und Trajektorien der Teilchen das komplette Event rekonstruiert, so dass die Zusammenhänge der Messwerte im Kontext des Experiments nachvollziehbar werden. Anschließend werden die Daten offline verarbeitet und der Datenanalyse zugeführt, welche die finalen Ergebnisse liefert.



Die CERN Datenpipeline [Abbildung 1]

2.3.2 CERN Cloud Controller

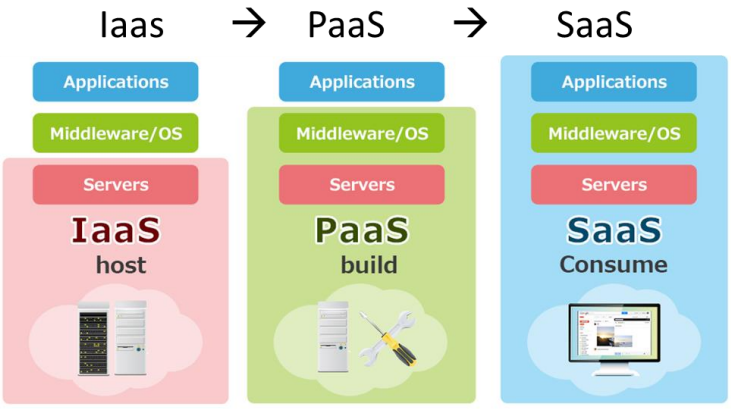
Der CERN Cloud Controller unterstützt sowohl Commercial als auch Open Source Cloud Umgebungen. Dadurch ist die CERN Cloud offen für viele Anforderungen. Beispiele für Cloud Environments die mit den virtuellen Maschinen im CERN kompatibel sind, können der folgenden Tabelle entnommen werden.

Hypervisor	**Cloud Controller**
Virtual Box	Vagrant
VMware	Open Stack
KVM	Open Nebula
Xen	Cloud Stack
Microsoft Hyper-V	Amazon EC2
	Google Compute Engine
	Micrrosoft Azure
	Docker

3 Cloud Computing in der Wissenschaft

3.1 Cloud Computing im Allgemeinen

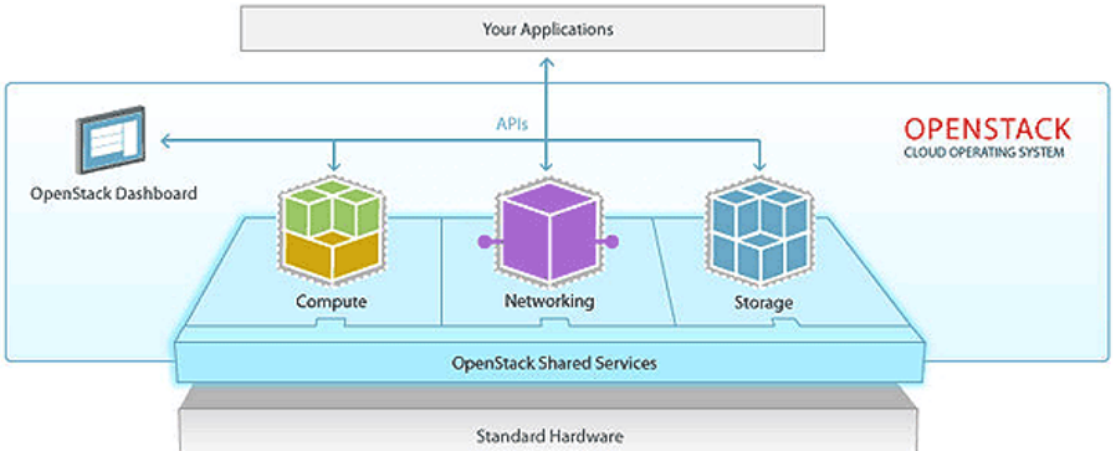
Die verbreitetsten Anwendungsgebiete von Cloud Diensten sind Infrastructure as a Sevice(IaaS), Plattform as a Service(PaaS) und Software as a Service(SaaS). (Abb. 4) IaaS bildet die unterste Ebene des Cloud Computing. Sie macht es möglich grundlegende IT Ressourcen wie Rechenleistung, Speicherplatz und Netzwerk bedarfsgerecht in Anspruch zu nehmen. Der Zugang zu dieser virtuellen Hardware erspart die eigene Anschaffung einer IT Infrastruktur. Beispiele für diese Ebene sind die Amazon Web Services S3 für Storage oder EC2 für Rechenleistung. Die Plattform as a Service dient als Verbindungsschicht zwischen IaaS und SaaS. Zusätzlich zur IaaS werden hier zum Beispiel Entwicklungsumgebungen, Web Server und Datenbank Verwaltungssysteme bereitgestellt, was PaaS besonders interessant für die Softwareentwicklung macht. Beispiele für diese Ebene sind Linux oder My SQL Datenbanken. Die SaaS ist die oberste Schicht des Cloud Computing und beschreibt vor allem die Anwendungssoftware der Cloudtechnologie. Der Zugriff wird meist über das Internet oder den Web Browser realisiert und die Nutzung erfolgt üblicherweise auf Abobasis. Dienste die sich in dieser Schicht eingliedern sind Drop Box oder die DESY Cloud.



Cloud Serviceklassen [Abbildung 4]

3.2 Open Stack

Open Stack ist ein hilfreiches Open Source Software Angebot zur Cloud Architektur in der Programmiersprache Python. In der Wissenschaft ist es zum Beispiel von CERN und der NASA in Benutzung. Über das Open Stack Dashboard kann man von unterschiedlichen Softwareangeboten gebrauch machen. Über APIs kann man die Bereiche Computing, Networking und den Storage verwalten. (Abb. 2) Inklusive Ad ONs kommt man auf 62 verschiedene Services im Angebot des Open Stack. Auch das Open Stack CERN wächst. Bis Ende 2015 ist die Anzahl der Open Stack Git Contributer ist auf über 700 pro Monat gestiegen und es spielt auch laut aktuellen Angaben auf der Website immernoch eine wichtige Rolle beim CERN.



Open Stack [Abbildung 2]

3.3 Metadaten

Metadaten spielen eine wichtige Rolle bei einer gut strukturierten Cloud. Es handelt sich um strukturierte Daten die Informationen über den Inhalt von anderen Daten enthalten. Sie beschreiben deren Merkmale. Metadaten werden benötigt für Standardisierung, also damit Daten vergleichbar sind mit anderen Daten, beispielsweise Daten aus anderen Experimenten oder Forschungsgruppen. Ebenfalls erleichtern sie die Auffindbarkeit von Daten. Sie ermöglichen einen schnellen Zugriff und erleichtern das Zusammenführen von Datensätzen, wenn neue Dateien gespeichert werden. Außerdem helfen Metadaten bei der Wiederverwendbarkeit und Nachvollziehbarkeit von Daten, weil durch sie verschiedene Forscherteams die Daten benutzen können ohne beispielsweise nachfragen zu müssen, welche Messgeräte benutzt wurden und unter welchen Rahmenbedingungen ein Experiment durchgeführt wurde. Beispiele für Metadaten die bei DESY Experimenten anfallen sind die Beschreibung des Experiments oder der Name des Beschleunigers welcher zum Einsatz kam. (Abb. 3)

Metadata common to all DESY experiments.

NeXus field or attribute path	ICAT field	Remark
/NXentry/experiment_identifier	Investigation	a unique beamtime ID
/NXentry/title	Investigation.Summary	description of the experiment
/NXentry/NXinstrument/name	Instrument.Fullname	name of the beamline
/NXentry/NXinstrument/name@short_name	Instrument.Name	short name of the beamline
/NXentry/NXinstrument/NXsource/name	Facility.Fullname	accelerator name

DESY Metadaten [Abbildung 3]

3.4 Anforderungen an Scientific Clouds

Eine wichtige Anforderung an Scientific Clouds ist eine garantierte Langzeitspeicherung der wissenschaftlichen Daten. Auf der Website des CERN äußern Forscher ihre Bedenken, dass große Commercial Cloud Anbieter wie Amazon oder Google eine Langzeitspeicherung in ihren Data Warehouses über Jahrzehnte garantieren können. "[...]*the Googles and Amazons of the world were building huge data warehouses to host commercial Clouds. Although we could turn to them to satisfy our computing needs, it is doubtful that such firms could guarantee the preservation of our data for the decades that it would be needed. We therefore need a dedicated "Science Cloud" instead*"(Eckhard Elsen, CERN) [32] Aus diesem Grund fördert das CERN die Arbeit an Scientific Clouds und wirkt an Projekten in diesem Bereich mit.

Die Entwicklung von Scientific Clouds benötigt sehr große Forschungsbudgets und stellt damit ebenfalls eine Herausforderung dar. Sie werden im Gegensatz zu Commercial Clouds nicht kommerziell betrieben und rentieren sich somit nicht finanziell.

Eine weitere Schwierigkeit für wissenschaftliche Clouds sind die großen Datenmengen welche oft in kurzer Zeit anfallen. Diese erfordern Technologien auf dem neusten Stand der Technik.

Cloud Dienste für die Wissenschaft werden oft für Individualentwicklungen verwendet, also für spezifische Zwecke, oftmals für ein einzelnes Projekt. Die eigentlichen

Anwendungen und Software Komponenten sind daher nur teilweise wiederverwendbar. Trotzdem entstehen auch hier viele Standardkomponenten und Bibliotheken, die in anderen Projekten wiederverwendet werden können. In der Regel sind dieses aber keine zusammenhängenden Softwaresysteme wie SAP, sondern modular aufgebaute Softwarebausteine, welche meist sehr spezifische Funktionen abbilden.

Ebenfalls wichtig beim Betrieb von Scientific Clouds ist es, einen Zugriff auf ältere Versionen zu haben, weil es bei wissenschaftlichen Arbeiten wichtig ist, dass die Ergebnisse reproduziert und somit getestet und überprüft werden können. Des Weiteren sollten häufige Updates, Anpassungen, sowie ein schnelles Deployment möglich sein. Es müssen also neue Versionen schnell und in häufiger kurzer Abfolge bereitgestellt werden.

4 Nutzen/Risiken

4.1 Nutzen

Da die notwendigen Investitionen zum Aufbau der Cloud nicht von den einzelnen Nutzern, sondern vom Cloud Provider getätigt werden und verursachungsgerecht auf die Nutzer verteilt werden, fallen nur Kosten für die tatsächliche Nutzung der Infrastruktur und nicht für deren Aufbau und Instandhaltung an. Dadurch wird auch finanzschwächeren Forschungsprojekten Zugang zu leistungsfähiger Technologie ermöglicht. Durch gemeinsame Nutzung der Ressourcen der Cloud kann diese sehr hoch ausgelastet werden und ist kostenmäßig verglichen mit einer individuell aufgebauten Infrastruktur, günstiger. Wissenschaftliche Organisationen müssen nun nicht mehr die gesamte Technologie, die für ihre Forschung benötigt wird, kaufen, sondern können bedarfsgerecht Rechenleistung in Anspruch nehmen. Leerkosten fallen damit weg, da nur noch Kosten in der Zeit entstehen in der die Rechner wirklich genutzt werden.

Ein weiterer Nutzen der Cloudtechnologie ist, dass auf Science Clouds Wissenschaftler von jedem beliebigen Ort und zu jeder beliebigen Zeit auf die wissenschaftlichen Daten zugreifen und mit ihnen Arbeiten können. Auch Kooperation bei internationalen Forschungsprojekten werden so vereinfacht oder freiwilligen Arbeit aus dem Home Office möglich gemacht. Bei CERN können sich Wissenschaftsinteressierte online auf der Website als Computing Volunteer registrieren und ihren Computer, während sie ihn nicht selbst nutzen, dem Cloud Netzwerk für zusätzliche Rechenleistung zur Verfügung stellen.

4.2 Risiken

Allerdings birgt die Cloud Technologie auch Risiken. Unter IT Sicherheitsexperten wird die Nutzung von Clouds skeptisch gesehen. Hackerangriffe sind keine Seltenheit und ein großes Problem, wenn es darum geht sensible Daten, die möglicherweise auch sehr wertvoll sein können, sicher online zu verwahren. Des Weiteren muss das Risiko einer Verletzung des Datenschutzes in Kauf genommen werden, falls man auf unternehmensexterne Clouds zurückgreift. Man darf beispielsweise nicht vernachlässigen, dass ein Großteil der Daten die auf Clouds gesichert sind, auf Rechnern in den USA gelagert werden. In anderen Ländern gelten andere Datenschutz Gesetze, weshalb man besonders Vorsichtig sein sollte, wenn es darum geht wem man seine Daten anvertraut.

Kritiker der voranschreitenden Digitalisierung beklagen den hohen Stromverbrauch von Rechenzentren. Allerdings lässt sich dem entgegen stellen, dass erneuerbare Energien

zur Stromversorgung genutzt werden und viel Optimierung im Bezug auf die Auslastung der Ressourcen betrieben wird.

Laut der Deutschen Forschungsgemeinschaft sollen alle Rohdaten zu wissenschaftlichen Publikationen mindestens für 10 Jahre zur Überprüfung und Nachvollziehbarkeit gespeichert werden. Deshalb stellen sich zumindest externe Commercial Cloud Anbieter als ungeeignet dar, da es zweifelhaft ist ob diese eine Langzeitspeicherung über Jahrzehnte garantieren können [30].

5 Ausblick

Es gibt einige Herausforderungen mit dem Einsatz von Cloud Technologie für wissenschaftliche Zwecke, wie zum Beispiel die Finanzierung von Scientific Clouds oder auch Unsicherheiten mit dem Datenschutz beim Zurückgreifen auf Commercial Clouds.

Die Potenziale, welche die Cloud Technologie birgt, rechtfertigen es jedoch sich den Herausforderungen, die sie mit sich bringt, zu stellen. Insbesondere die Möglichkeit IT Ressourcen bedarfsgerecht in Anspruch zu nehmen, ohne sich selber die nötige Hardware anzuschaffen, ist ein großer Vorteil. Sie hilft den Wissenschaftlern im Umgang mit der Datenflut, welche durch voranschreitende Technologien, wie zum Beispiel leistungsstärkere Teilchenbeschleuniger, exponentiell wächst. Das Arbeiten von Wissenschaftlern an gemeinsamen Projekten kann durch einen orts- und zeitunabhängigen Zugriff auf ihre Daten, erleichtert werden. Außerdem bieten sich neue, experimentelle Möglichkeiten Freiwillige in die Forschung miteinzubeziehen. Diese können beispielsweise ihre Computer ins Cloud Netzwerk integrieren.

Da Clouds großen Nutzen für die Wissenschaft bringen können und sie die Forschungsprozesse effizienter machen, ist anzunehmen, dass Clouds bald standardmäßig zum Einsatz bei großen Forschungsprojekten und in Großforschungseinrichtungen kommen werden.

Abbildungen

[Abbildung 1] https://image.slidesharecdn.com/01enrictejedor-181010193440/95/cerns-next-generation-data-analysis-platform-with-apache-spark-with-enric-tejedor-9-638.jpg?cb=1539200135

[Abbildung 2] http://www.sdnstack.com/wp-content/uploads/2015/07/openstack-sm.png

[Abbildung 3] http://www.desy.de/ jkotan/nexusdesyintegration.pdf S.22

[Abbildung 4] https://pbxl.co.jp/wordpress/wp-content/uploads/2015/05/cloud_-en.png

Quellen

[1] https://home.cern/science/computing

[2] https://home.cern/science/computing/volunteer-computing

[3] https://home.cern/news/news/cern/european-science-cloud-open-day

[4] https://home.cern/science/computing/data-centre

[5] https://home.cern/science/computing/storage

[6] http://eos.web.cern.ch/

[7] https://www.youtube.com/watch?v=qIn0homyLm4

[8] https://www.bmbf.de/de/auf-dem-weg-zur-european-open-science-cloud-5216.html

[9] https://sureshemre.wordpress.com/2015/02/22/cern-lhc-large-hadron-collider-is-waking-up/

[10] https://www.golem.de/news/wissenschaft-10-000-cpus-und-1-petabyte-fuer-die-cloud-von-helix-nebula-1802-132802.html

[11] https://www.cloudcomputing-insider.de/metadaten-sind-das-wertvollste-gut-in-einer-multicloud-welt-a-782375/

[12] https://de.ryte.com/wiki/Metadaten

[13] https://phys.org/news/2018-06-cern-major-reap-atom-smasher.html

[14] https://en.wikipedia.org/wiki/OpenNebula

[15] https://de.wikipedia.org/wiki/CERN

[16] https://www.abendblatt.de/hamburg/article213327465/Roentgenlaser-700-Terabyte-Daten-fuer-340-Forscher.html

[17] https://cloud.google.com/blog/topics/inside-google-cloud/subatomic-particles-and-big-data-google-joins-cern-openlab

[18] https://www.youtube.com/watch?v=N4eT9Lfvuro

[19] http://www.desy.de/ jkotan/nexusdesyintegration.pdf

[20] https://www.slideshare.net/databricks/cerns-next-generation-data-analysis-platform-with-apache-spark-with-enric-tejedor

[21] https://www.datenschutzbeauftragter-info.de/die-cloud-saas-paas-und-iaas-einfach-erklaert/

[22] https://www.mpg.de/6957705/JB_2013

[23] http://www.desy.de/forschung/beschleuniger/index_ger.html

[24] https://it.desy.de/dienste/speicherdienste/desycloud/index_-ger.html

[25] https://de.wikipedia.org/wiki/Cloud_ComputingCloud_-Computing_in_der_Wirtschaft

[26] http://www.desy.de/femto/sites2009/site_www-desy/content/e187923/e187955/e198864/femto_2015_-1_dt_ger.pdf

[27] http://www.desy.de/aktuelles/news_suche/index_-ger.html?openDirectAnchor=1213two_columns=1

[28] https://cernvm.cern.ch/sites/cernvm.web.cern.ch/files/hep-cloud.pdf

[29] https://www.weltderphysik.de/gebiet/teilchen/experimente/teilchenbeschleuniger/lhc/lhc-faq/

[30] https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

[31] http://www.desy.de/forschung/anlagen___projekte/tier_-2/index_ger.html

[32] Eckhard Elsen, 12.08.2016, "Viewpoint: The end of computing's steam age", https://cerncourier.com/viewpoint-the-end-of-computings-steam-age/ [Stand 24.02.2019]