

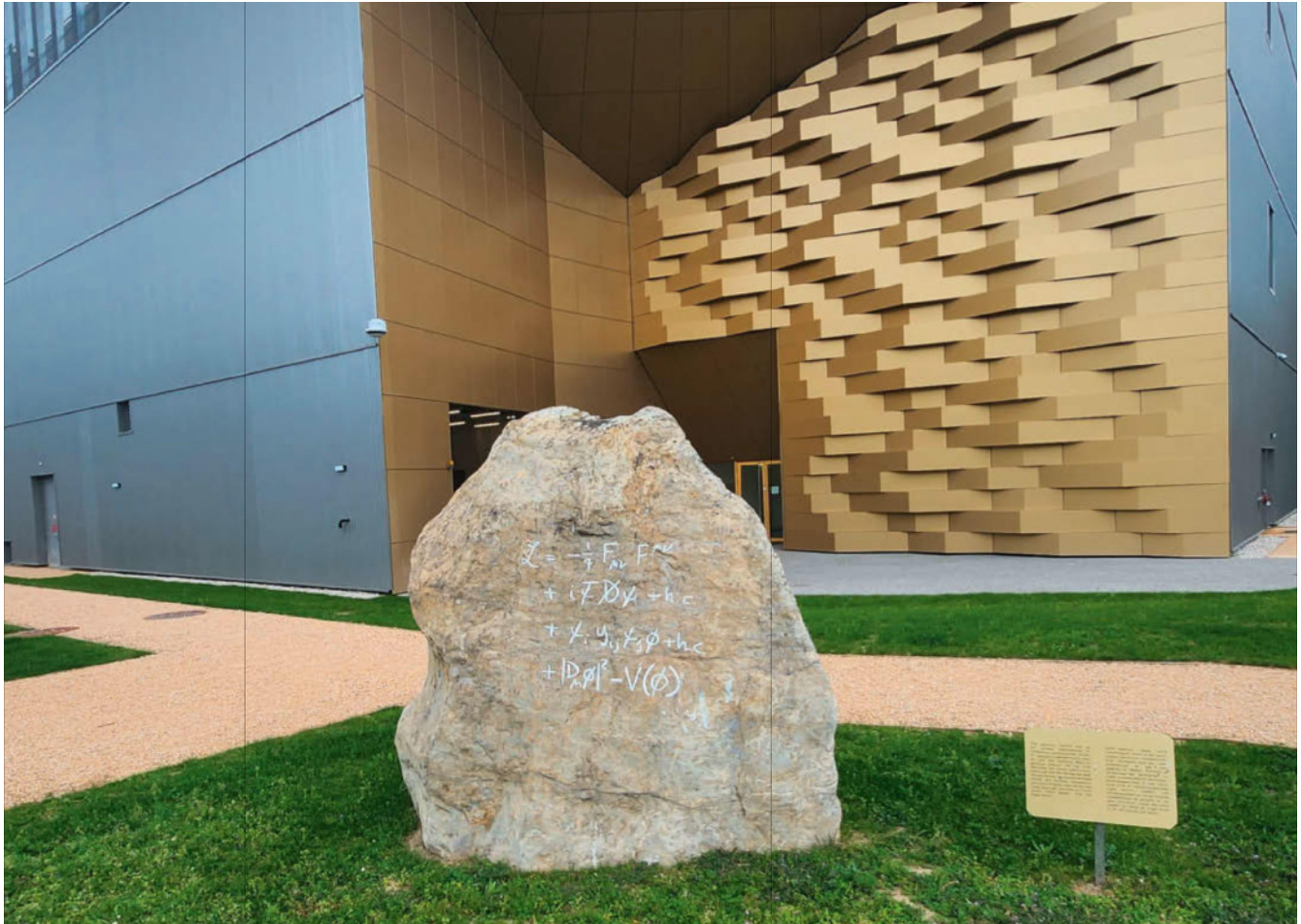


Bild: Andreas Peters

Massenhaft Massenspeicher

Die Storage-Architektur des Kernforschungszentrums CERN

Der größte Teilchenbeschleuniger der Welt produziert in jeder Sekunde Datenmengen im zweistelligen Terabyte-Bereich. Nur ein kleiner Teil dient der Forschung. Um diese Datensturmflut einzudämmen, kombiniert das CERN Eigenentwicklungen und Open Source.

Von Andreas Peters

Das weltweit größte Labor für Teilchenphysik produziert Datenmengen, die konventionelle IT-Infrastrukturen an ihre

Grenzen bringen. Im Folgenden schildern wir, wie das CERN mit Open Source, innovativen Speicherarchitekturen und globaler Zusammenarbeit das Datenmanagement des Large Hadron Collider (LHC) meistert und welche Herausforderungen die Zukunft mit sich bringt.

Der LHC ist ein 27 Kilometer langer Ringbeschleuniger in einem Tunnel rund 100 Meter unter der Erde an der Grenze zwischen Frankreich und der Schweiz. Er hat die Aufgabe, Teilchen bei bisher unerreichten Energien kollidieren zu lassen, um die fundamentalen Bausteine der Materie und die Kräfte zwischen ihnen zu erforschen, etwa das Higgs-Boson, Dunkle Materie und die Asymmetrie zwischen Materie und Antimaterie. Wer sich dem CERN-Besucherzentrum nähert, be-

kommt einen Vorgeschmack auf diese theoretischen Grundlagen: Auf dem Stein davor ist die Lagrange-Dichte des Standardmodells der Teilchenphysik eingraviert (siehe Aufmacherbild).

Der Betrieb läuft in mehreren Phasen, sogenannten Runs: Run 1 (2009 bis 2013) führte zur Entdeckung des Higgs-Bosons, Run 2 von 2015 bis 2018 steigerte Energie und Kollisionsrate, und der beim Erscheinen dieses Artikels gerade abgeschlossene Run 3 (seit 2022) arbeitete mit noch höherer Luminosität (Intensität oder Dichte des Teilchenstrahls) und lieferte durch die höhere Anzahl der Teilchenkollisionen noch einmal deutlich mehr Daten als die Vorgängerläufe.

Das Speichern aller vom LHC erzeugten Rohdaten aber ist weder möglich noch

notwendig. Bei Proton-Proton-Kollisionen am LHC kommt es zu etwa ein bis zwei Milliarden Kollisionen pro Sekunde. Jedes Ereignis erzeugt mehrere Megabyte an Daten – eine gigantische Rohdatenrate, die die Speicherkapazität aller aktuellen Systeme übersteigt. Um die Informationsflut zu bändigen, greifen mehrstufige Triggersysteme ein: Sie filtern aus allen Kollisionen nur die wissenschaftlich interessantesten heraus. Ein Triggersystem ist ein Echtzeitfilter, der Kollisionen anhand charakteristischer Messsignale auswählt, beispielsweise besonders energiereicher Teilchen oder ungewöhnlicher Teilchenmuster und nur potenziell interessante Ereignisse zur weiteren Analyse durchlässt.

Viel Open-Source

Das CERN setzt seit Jahren konsequent auf Open-Source-Lösungen für den Massenspeicher, teils selbst entwickelte. Dies ermöglicht dem Forschungszentrum, gezielt auf die Anforderungen der Experimente einzugehen und Innovationen gemeinsam mit der Open-Source-Community voranzutreiben.

CERN Tape Archive (CTA) ist das Magnetbandspeichersystem zur langfristigen und kosteneffizienten Archivierung wissenschaftlicher Daten. Es verwaltet weit mehr als ein Exabyte an Daten auf Bandmedien und archiviert Forschungsdaten über Jahrzehnte hinweg, damit sie bei Bedarf wieder zur Verfügung stehen. CTA wurde als Open-Source-Nachfolger früherer Archivsysteme entwickelt und ist auf höchste Skalierbarkeit und Effizienz ausgelegt.

EOS Open Storage ist das hochperformante, verteilte Festplattenspeichersystem des CERN und wurde dort speziell für die Anforderungen der Hochenergiephysik entwickelt. Es bietet extrem hohe Datendurchsätze, eine skalierbare Namespace-Verwaltung (Katalog aller Pfade und Dateinamen) und einen effizienten Zugriff auf Milliarden von Dateien. Dadurch eignet sich EOS besonders zur Speicherung und Analyse der großen Datenmengen, die die Experimente erzeugen.

Dazu gesellt sich Ceph, ein verteiltes Objektspeicher-System. Ceph spielt am CERN eine wichtige Rolle als skalierbare und ausfallsichere Plattform für Objekt-, Block- und Cloud-native Storage-Anwendungen. Als Backend für EOS dient Ceph jedoch nicht, denn für die spezifischen Workloads im Exabyte-Maßstab verur-

ct kompakt

- CERN setzt auf Open-Source-Speicherlösungen und entwickelt eigene Systeme speziell für seine Anforderungen.
- Die Speicherhierarchie aus DRAM, SSD, Festplatte und Magnetband verbessert Effizienz, Flexibilität und Wirtschaftlichkeit.
- Mit dem HL-LHC und dem geplanten FCC drohen künftig nochmals erheblich höhere Datenraten sowie neue technische Herausforderungen.

sacht Ceph höhere Betriebs- und Infrastrukturkosten als die eigens entwickelte EOS-Plattform.

Mehrstufiges Speichermode

Direkt an die LHC-Detektoren angebundene DRAM- und SSD-Puffer erfassen die kontinuierlich eintreffenden Datenströme. Auf dieser Ebene steht die maximale Schreibgeschwindigkeit im Vordergrund, um Datenverluste zu vermeiden und die Echtzeitverarbeitung zu gewährleisten.

Beim ALICE-Experiment etwa lesen First-Level-Prozessoren (FLP) den Datenstrom mithilfe von FPGAs mit einer Rohdatenrate von etwa 28 Tbit/s aus. Sie bündeln die Daten zahlreicher optischer Fasern und reduzieren den Strom, indem sie

irrelevante Daten und leere Datenpakete löschen; rund 20 Prozent der Daten bleiben dabei übrig. Anschließend übertragen sie bis zu 5,4 Tbit/s (679 GByte/s) an eine GPU-Farm. ALICE erfasst die Daten kontinuierlich in 20 Millisekunden langen Zeitscheiben. Während andere Experimente einzelne Ereignisse per Trigger auswählen und auslesen, verarbeitet ALICE den gesamten Datenstrom fortlaufend, ähnlich wie eine Filmkamera mit konstanter Bildrate.

Das Storage-System EOS ist der zentrale Physikdatenspeicher und stellt derzeit etwa 1,16 Exabyte an Festplattenspeicherplatz zur Verfügung, verteilt auf Dutzende Instanzen. Die Datentransferrate des Gesamtsystems liegt bei etwa 1 TByte/s, insgesamt liegen rund 8,7 Milliarden Dateien auf den Servern. Die größte EOS-Instanz für das ALICE-Experiment speichert 150 PByte an Daten im Erasure-Coding-Format 10+2 mit Dual Parity und 20 Prozent Volumen-Overhead. Das ist eine technische Wahl mit einem Kompromiss zwischen tolerierbarem „Repair Traffic“ und „Volumen-Overhead“ und hat sich im Betrieb gut bewährt. Es gibt jedoch Überlegungen, aus Kostengründen mit weniger Volumen-Overhead zu experimentieren, etwa dem Erasure-Coding-Format 16+2.

Durch die recht große Sicherheit vor Datenverlusten erlaubt das CERN eine Betriebsdauer von bis zu zehn Jahren pro Laufwerk, doppelt so lang wie der für Enterprise-Laufwerke übliche Garantiezeitraum der Hersteller.

Zur Archivierung wandern die Daten auf Bänder: Tape Libraries bewahren rie-



In zwei Rackreihen stecken 50 Petabyte, ein Drittel des gesamten Speichervolumens EOS ALICE O2 Storage Systems im Meyrin-Rechenzentrum.

sige Datenmengen kostengünstig und energieeffizient über Jahrzehnte hinweg auf. Das CERN rechnet für die Archivierung mit etwa 10 bis 25 Prozent der Kosten, die für ein Festplattensystem aufzubringen wären. Etwa zwei Drittel der Kosten für die Archivierung entfallen auf den Kaufpreis der Bänder, der Rest auf die Infrastruktur. Die Tape-Kapazität liegt aktuell bei etwa 1,7 Exabyte, das Datenvolumen allein aus dem Run 3 bei rund 1 Exabyte.

Der Ceph-Speicher hat eine vergleichsweise kleine Kapazität. Außer etwa 133 Petabyte Festplattenplatz bietet das System rund 10 Petabyte SSD-Speicher für schnelleren Zugriff. Ceph wird insbesondere als Backend für Infrastruktur-Services mit hohen Verfügbarkeitsanforderungen benutzt. Darunter fallen etwa die CERN-Homepage oder virtuelle Maschinen mit OpenStack. Dazu dient Ceph als Lager für Object Storage für Gitlab, Kubernetes und Openshift.

Jede Ebene dieser Hierarchie erfüllt eine spezifische Aufgabe: DRAM- und SSD-Puffer für Geschwindigkeit, Festplatten für Analyse und schnellen Zugriff, Tape für langfristige Sicherheit.

Zudem spiegelt das CERN die Daten in das Worldwide LHC Computing Grid (WLCG). Diese weltweit verteilte Recheninfrastruktur betreibt das CERN zusammen mit über 170 Rechenzentren in mehr als 40 Ländern, um die enormen Datenmengen des Large Hadron Collider zu speichern, zu verteilen und zu analysieren. Tausende von Physikern nutzen es für Simulationen und Analysen. Das WLCG kombiniert rund 1,3 Millionen Rechenker-

ne und über 4 Exabyte Speicher, bewältigt Millionen von Rechenjobs pro Tag und erreicht Datentransferraten von weit über 200 GByte/s. Dabei nutzt das WLCG ein Tiering-Modell, um die Daten möglichst effizient zu verteilen: Tier 0 liegt innerhalb des CERN, Tier 1 sind nationale Zentren und Tier 2 daran angeschlossene Universitäten.

Metadaten, Dateiformate und Langzeitarchivierung

Daten sind nur dann nutzbar, wenn sie auffindbar und korrekt interpretierbar bleiben. Daher dokumentiert das CERN die Messdaten mit umfangreichen Metadaten, etwa zu Detektorstatus, Zeitstempeln und Triggerbedingungen. Die Speicherung geschieht häufig im ROOT-Format, das am CERN speziell für die Verarbeitung großer Datenmengen in der Teilchenphysik entwickelt wurde.

Aktuell steht die Einführung von RNTuple an, einem neuen spaltenbasierten Datenformat im ROOT-Ökosystem, das für hochparallele Analysen und die effiziente Nutzung moderner Storage-Systeme entwickelt wurde. RNTuple arbeitet mit Transformationsbibliotheken und modernen Kompressionsbibliotheken wie ZSTD, was sowohl die Performance als auch den Speicherbedarf weiter optimiert. In vielen Anwendungsfällen kommt RNTuple mit 40 bis 50 Prozent weniger Speicherplatz aus als das TTree ROOT-Format.

Langzeitarchivierung

Ein negativer Aspekt bei der Archivierung auf Tape ist die mittlerweile eingeschränk-

te Rückwärtskompatibilität des LTO-Standards. Die langjährige Regel „Lese zwei Generationen zurück, schreibe eine Generation zurück“ wurde ab LTO-9 eingeschränkt, die aktuellen LTO-10-Laufwerke können gar keine Bänder früherer Generationen mehr lesen. Für viele Tape-Storage-Betreiber bedeutet das außer der Anschaffung neuer Laufwerke und Medien den Aufbau einer zweiten Tape-Infrastruktur für den Übergang zur aktuellen Version.

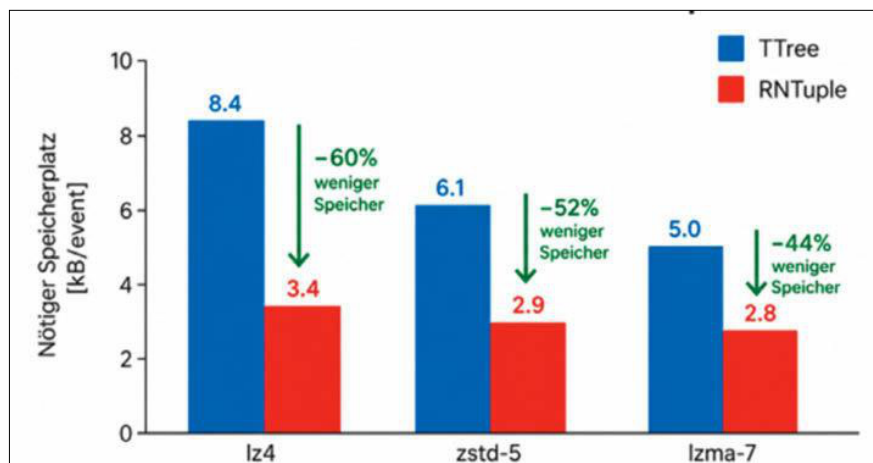
Die Sicherung unersetzbarer Forschungsdaten über Jahrzehnte hinweg erfordert zudem regelmäßige Medienmigration, Prüfsummen, Redundanz und sorgfältige Dokumentation – allein um die Lesbarkeit in künftigen Systemen zu gewährleisten. Zudem sollten die Daten redundant an verschiedenen Standorten lagern.

Das CERN erprobt regelmäßig neue Speicher- und Archivierungstechniken in Zusammenarbeit mit Industriepartnern wie IBM, Siemens und Everpure sowie mit innovativen Start-ups wie dem deutschen Unternehmen Cerabyte [1]. Im Rahmen von CERN openlab arbeitet das CERN Storage-Team gemeinsam mit Cerabyte daran, die Entwicklung eines keramikbasierten Systems für die Langzeitarchivierung voranzutreiben.

Nachhaltigkeit trotz steigender Anforderungen

Der Energiebedarf der Speichersysteme ist im Vergleich zu den Beschleunigeranlagen gering. Tape etwa benötigt nur beim Zugriff Strom; die meisten Tapes lagern jedoch in einer definierten klimatisierten Umgebung. Festplatten, vor allem in großen Mengen, sparen durch gezieltes Herunterfahren [2] viel Energie; das CERN identifiziert gerade zusammen mit einem Partner, wie sich das für JBODs realisieren lässt.

Das neue CERN-Rechenzentrum in Prévessin setzt für eine nachhaltige IT-Infrastruktur auf Wärmerückgewinnung und niedrige Power Usage Effectiveness (PUE). Die PUE setzt die Gesamtleistungsaufnahme des Rechenzentrums inklusive Beleuchtung und Kühlung in Relation zur elektrischen Leistung, welche die eigentliche IT-Infrastruktur umsetzt, beispielsweise die Server. Der PUE-Zielwert für Prévessin liegt bei niedrigen 1,1, das Rechenzentrum nimmt also nur 10 Prozent mehr Leistung auf, als die IT-Geräte benötigen.



Ein anderer Kompressionsalgorithmus und der Wechsel vom TTree- zum RNTuple-Format reduzieren den notwendigen Speicherplatz für die LHC-Daten.

Bild: CERN

Das Wärmerückgewinnungssystem soll ab dem kommenden Winter mit der Abwärme des Rechenzentrums die Gebäude am Standort heizen und damit einen wichtigen Beitrag zur Steigerung der Energieeffizienz leisten. Der Großteil von CERNs Storage aber steht derzeit im Meyrin-Rechenzentrum auf dem Hauptcampus.

HL-LHC, FCC und die Zukunft der Speicherinfrastruktur

Run 3 des LHC ist beendet, aktuell befindet sich der Teilchenbeschleuniger in einer Umbauphase, Long Shutdown 3 oder kurz LS 3 genannt. Mit der geplanten Inbetriebnahme des High-Luminosity LHC (HL-LHC, Runs 4 und 5) um das Jahr 2030 werden Datenmengen von bis zu einem Exabyte pro Jahr erwartet. Die größten Engpässe erwarten die Forscher bei den I/O-Raten, der Langzeitarchivierung und der Grid-Bandbreite.

CERN investiert dafür bereits in neue Tape-Techniken und automatisierte Archivsysteme, hybride Speicherarchitekturen aus SSDs, Festplatten und Tapes. Machine-Learning-Algorithmen identifizieren Daten auf Festplatten, die lange nicht benutzt wurden und daher auf das günstigere Tape verschoben werden können. Die zentralen Herausforderungen bestehen in der Entwicklung skalierbarer Speicherlösungen, der weiteren Automatisierung des Datenmanagements und einer hohen Energieeffizienz.

Die durch den KI-Boom stark gestiegenen Preise für Festplatten und Storage-



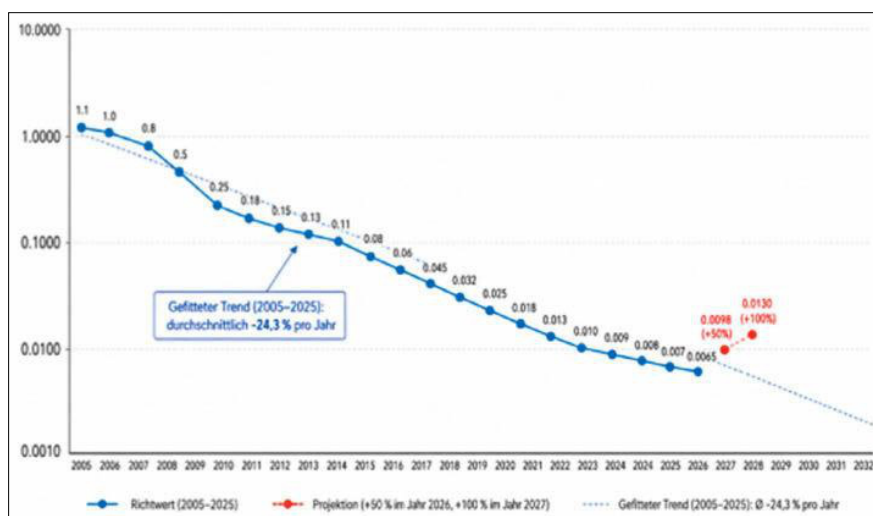
Das auf Tape gespeicherte Datenvolumen der LHC-Experimente von 2012 bis 2026 über die verschiedenen Betriebsphasen ist hier logarithmisch dargestellt, real verläuft es exponentiell.

Hardware erschweren auch den Ausbau wissenschaftlicher Infrastrukturen. Über viele Jahre galt für Storage ein abgewandeltes Moore'sches Gesetz mit einer jährlichen Steigerung der Kapazität von mindestens 20 Prozent bei gleichen Kosten. Seit Anfang 2026 aber stiegen die Preise für Festplatten in den CERN-Hardware-Ausschreibungen um rund 50 Prozent, für 2027 sind weitere Preissteigerungen zu erwarten.

Das aktuelle LHC-Programm einschließlich des HL-LHC soll bis in die 2040er-Jahre hinein laufen. Danach könn-

te der Future Circular Collider (FCC) die Forschung am CERN fortführen, ein Teilchenbeschleuniger der nächsten Generation. Diesem Ausbau müssen jedoch die CERN-Mitgliedsländer zustimmen. Die erste Ausbaustufe FCC-ee lässt bei einem klassischen Trigger-Readout rund 40 GByte/s erwarten, also ähnliche Datenmengen wie der HL-LHC. Aber es gibt Überlegungen, einen Streaming-Readout mit etwa 100 GByte/s oder sogar einen vollständig ungefilterten Readout mit bis zu 10 TByte/s einzusetzen.

Darüber hinaus könnten die Datenraten des künftigen FCC-hh die heutigen Anforderungen bei Weitem übersteigen und bis zu 250 TByte/s für Kalorimeter- und Myon-Systeme sowie bis zu 2 PByte/s für Tracker erreichen. Solche Datenmengen dürften erhebliche technische Herausforderungen mit sich bringen. Die am CERN entwickelten quelloffenen Hochleistungsspeichersysteme sind jedoch gut aufgestellt, um diesen Herausforderungen zu begegnen. Ihre zunehmende Nutzung über die Teilchenphysik hinaus stärkt zudem ihren Beitrag zur europäischen Wissenschafts- und Technologie-Landschaft. (ll@ct.de)



Der Preis pro Speichervolumen für die Storage-Hardware ist in den vergangenen Jahren stetig gesunken. Nun aber sind durch den KI-Bauboom extreme Steigerungen zu befürchten.

Literatur

- [1] Lutz Labs, Steintafel en miniature, Langzeitarchivierung auf Glasplättchen mit Keramikbeschichtung und Femtosekundenlaser, c't 1/2026, S. 124
- [2] Lutz Labs, Intelligent abschalten, Stromsparen bei großen Festplatten-Arrays, c't 7/2026, S. 118