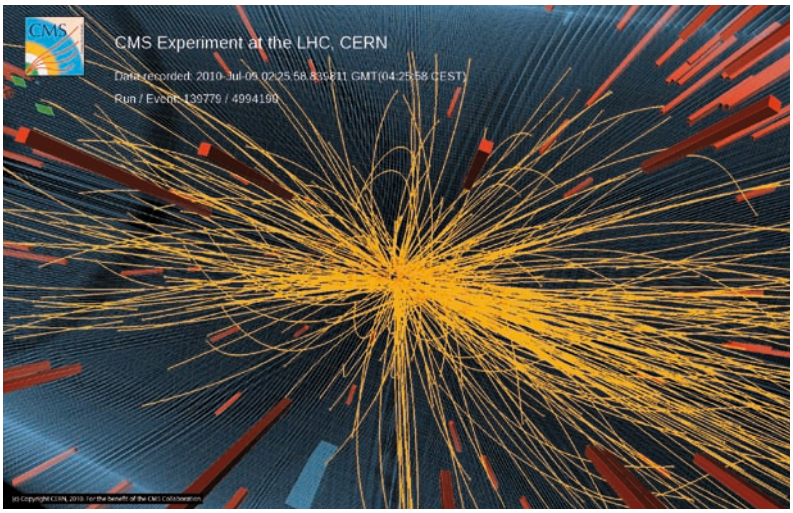
TEILCHENPHYSIK

Rechnen im Netz

Das Grid Computing ist für die Datenanalyse der LHC-Experimente unentbehrlich.

Günter Quast und Armin Scheurer

Mit dem Start des regulären Betriebs des Large Hadron Collider (LHC) am CERN begann für die Teilchenphysik eine neue Ära. In dieser sollen sich zentrale Fragen klären wie die nach dem Ursprung der Masse oder nach der theoretisch vermuteten Supersymmetrie zwischen Fermionen und Bosonen. Die beteiligten Physiker haben sich aber auch auf vielerlei Szenarien neuer Physik jenseits des bisher äußerst erfolgreichen Standardmodells der Teilchenphysik vorbereitet. Das Rückgrat für die Datenauswertung der Experimente bildet ein globales Netzwerk von mehreren hundert Rechenzentren, das „Worldwide LHC Computing Grid".



Bei einer Proton-Proton-Kollision am LHC können hunderte Sekundärteilchen entstehen wie bei diesem von CMS gemessenen Ereignis. Damit einher geht eine enorme Datenflut, die ein weltweit verteiltes Computernetz, das LHC-Grid, bewältigen soll.

Nach jahrzehntelangen Aufbauarbeiten des LHC und der Detektoren ALICE, ATLAS, CMS und LHCb konnten am 10. September 2008 alle Experimente erstmals Strahlreaktionen der Protonen im LHC bei einer Injektionsenergie von 450 GeV aufzeichnen. Während der mehr als einjährigen Reparaturphase nach einer technische Panne im September 2008 gelang es, Milliarden Ereignisse aus der kosmischen Strahlung zu registrieren und damit eine erste Eichung der Detektorkomponenten vorzunehmen. Die Zeit diente insbesondere auch dazu, die Verteilung und Auswertung der Daten innerhalb des Computer-Netzwerks der am LHC beteiligten Institute unter realistischen Bedingungen zu erproben und zu verbessern. Bei der Wiederinbetriebnahme Ende 2009 löste der LHC mit dem Erreichen einer Schwerpunktsenergie von 2,36 TeV schließlich das Tevatron als weltweit leistungsstärksten Beschleuniger ab. Nach einer kurzen Winterpause lief der LHC dann ab dem 30. März 2010 im regulären Betrieb bei einer Schwerpunktsenergie von 7 TeV.

Alle vier Experimente haben die in den Kollisionen entstandenen Ereignisse mit hoher Effizienz aufgezeichnet. Die schon im Sommer 2010 auf der International Conference on High Energy Physics in Paris veröffentlichten ersten Ergebnisse belegen das hervorragende Funktionieren aller Detektoren [1]. Mittlerweile haben die LHC-Experimente die bisher bekannten Teilchen des Standardmodells nachgewiesen und ihre Produktionsraten in Kollisionen bei 7 TeV gemessen.

Während der gesamten Datennahme wurde die Kollisionsrate im LHC kontinuierlich gesteigert. Insgesamt lieferte der LHC im ersten Jahr eine integrierte Luminosität von etwa 45/pb – das entspricht rund einer halben Million produzierter W-Bosonen, 50 000 Z-Bosonen oder 1000 Paaren von Top-Quarks. In diesem Jahr soll die Anzahl der Protonenpakete im Strahl nochmals deutlich ansteigen, sodass Ende des Jahres 2011 mit einer integrierten Luminosität von mindestens 1/fb zu rechnen ist. Damit sind die Experimente und der LHC bestens für die Entdeckung neuer Physik gerüstet.

Von entscheidender Bedeutung für die rasche und erfolgreiche Datenauswertung in allen an den LHC-Experimenten beteiligten Instituten ist das globale Netzwerk von Rechenzentren, das in den letzten zehn Jahren geplant und aufgebaut worden ist. Der bei weitem größte Teil der dafür notwendigen Computing-Ressourcen befindet sich nicht am CERN selbst, sondern wird von einer Vielzahl von global verteilten Rechenzentren bereit gestellt.

KOMPAKT

- Die hierarchisch aufgebaute Computing-Infrastruktur am LHC erlaubt es, dezentral organisierte Rechner- und Speicherressourcen mittels standardisierter, offener Protokolle und Schnittstellen zu vernetzen.
- Dieses LHC Computing Grid hat die Erwartungen der über 8000 an den Experimenten beteiligten Physiker mehr als erfüllt und sich in nahezu allen Aspekten der Datenauswertung bewährt.
- Angesichts der von den vier Experimenten erwarteten Datenflut 2011/2012 gilt es nun, die vorhandenen Ressourcen optimiert zu nutzen.

Prof. Dr. Günter Quast, Dr. Armin Scheurer, Institut für Experimentelle Kernphysik, KIT, Wolfgang-Gaede-Str. 1, 76131 Karlsruhe

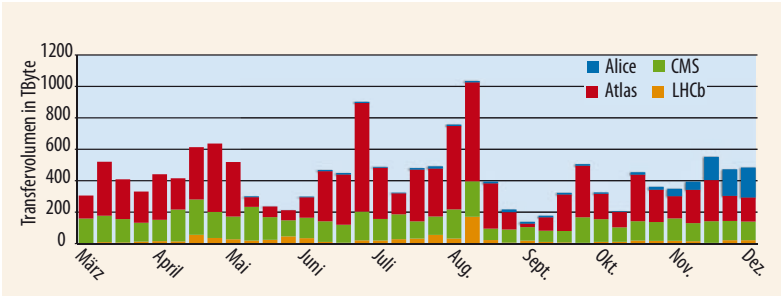


Abb. 1 Seit März 2010 wurden vom CERN aus Woche für Woche bis zu 1000 TByte an Daten übertragen.

Schon während der Anfangsphase der LHC-Datennahme musste das Computer-Netzwerk eine hohe Datenrate bewältigen, da die Experimente die für die Aufzeichnung vorgesehenen Ereignisse zunächst mit sehr losen Kriterien auswählten. Mit steigender LHC-Luminosität wurde der aufgezeichnete Datenstrom durch schärfere Vorselektion in den Trigger-Stufen praktisch konstant gehalten. Der Datentransfer vom CERN schwankte abhängig von den verwendeten Trigger-Bedingungen der Experimente, befand sich aber kontinuierlich auf einem sehr hohen Niveau und erreichte im August 2010 ein Maximum von über tausend Terabytes oder 1 Petabyte (PByte) pro Woche (**Abb. 1**) – das entspricht etwa 200 000 einlagigen DVDs, die gestapelt 200 Meter hoch wären. Diesen Datenstrom erwarteten die Teilchenphysiker – darunter rund tausend Doktoranden in jedem der beiden großen Experimente ATLAS und CMS – begierig, um ihn zu analysieren.

Die Suche im Heuhaufen

Die Herausforderung bei der Datenanalyse besteht darin, die interessanten, meist äußerst seltenen Ereignisse aus einer um viele Größenordnungen höheren Anzahl von weniger interessanten Kollisionen auszuwählen. Bei der Design-Luminosität des LHC kommt es pro Sekunde zu 40 Millionen Zusammenstößen der gegenläufig zirkulierenden Protonenpakete, und bei jeder dieser Kollisionen entstehen mehr als 1000 neue Teilchen, also insgesamt etwa 10^{11} Teilchen pro Sekunde. Die typischerweise mehr als 100 Millionen Detektorzellen liefern dabei eine Datenmenge in der Größenordnung von einer Million GByte/s, die sich jedoch durch Unterdrückung der Daten von nicht angesprochenen Zellen auf einige Tausend GByte/s reduzieren lässt. Um diesen dennoch enormen Informationsfluss zu bewältigen, muss bereits die erste Stufe der Datenselektion „online" ein Ereignis aus 100 000 auswählen. Das dazu notwendige Trigger-System besteht zum einen aus hoch-parallelisiert arbeitenden Prozessoren, die kommerziell nicht erhältlich sind und daher von Teilchenphysikern selbst entwickelt und gebaut werden [2], und zum anderen aus einer PC-Farm, deren Selektions-Algorithmen in Software realisiert sind.

Nur einige hundert Ereignisse pro Sekunde überstehen diese Selektion und werden zum einen an die „Offline"-Datenspeicherung und gleichzeitig auch zu einer ersten vollen Ereignis-Rekonstruktion weitergegeben. Dabei fällt immer noch eine Gesamtdatenmenge von mehreren PByte pro Jahr und Experiment für die detaillierte Analyse an Universitäten und Forschungsinstituten in der ganzen Welt an. Nach einer Auswahl von 1 aus bis zu 10^{12} der ursprünglich im Detektor registrierten Ereignisse gilt es schließlich, Signaturen zu finden, die neue physikalische Phänomene darstellen könnten.

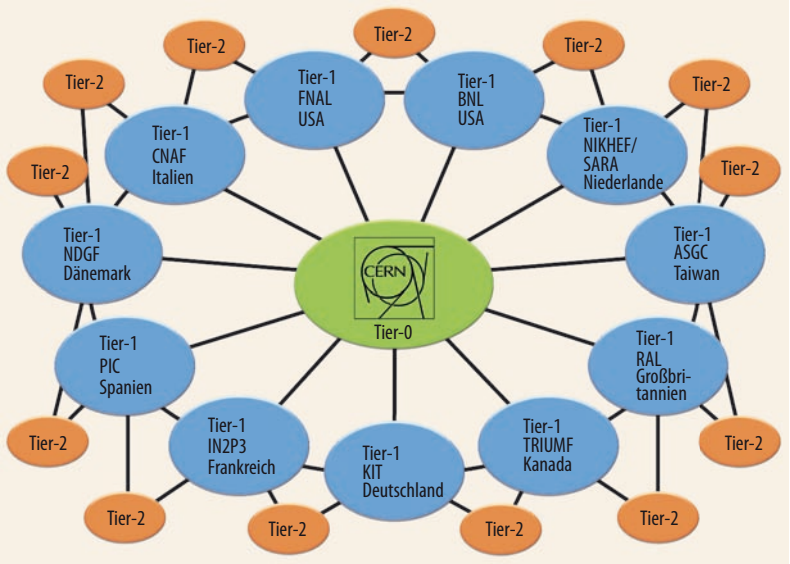


Abb. 2 Im Mittelpunkt des „Worldwide LHC Computing Grid" steht das CERN, das mit elf Zentren der Ebene 1 (Tier-1) und etwa 140 weiteren Standorten (Tier-2) vernetzt ist.

Jederzeit und überall bereit

Dem 1998 erstmals von Ian Foster und Carl Kesselmann definierten Grid-Computing liegt der Gedanke zugrunde, dass über ein weltumspannendes Netzwerk überall und jederzeit Rechenleistung zur Verfügung stehen soll, indem dezentral organisierte Rechner- und Speicher-Ressourcen mittels standardisierter, offener und allgemeiner Protokolle und Schnittstellen vernetzt werden [3]. Dabei ermöglicht die Grid-Infrastruktur, die Ressourcen koordiniert zu nutzen, um so Aufgaben zu lösen, die sich in dynamischen, verteilten Organisationen stellen – wenn etwa Wissenschaftler in Forschungszentren oder Mitarbeiter in multinationalen Unternehmen an verschiedenen Standorten gemeinsam an einer Aufgabe arbeiten und ihre jeweils lokalen Ressourcen in eine gemeinsam genutzte Infrastruktur einbringen. Als Ressource aufzufassen sind dabei nicht nur reine Rechenleistung, sondern ebenso Datenspeicher, Visualisierungssysteme, Infrastruktur und Bandbreite der Netzwerke, Spezialrechner zur Steuerung und Administration sowie Datenquellen, die nicht an jedem Ort verfügbar sind. In diesem Sinne sind auch die Datennahmesysteme der Experimente der Teilchenphysik selbst als besonderer Teil eines Computing-Grids zu sehen.

Die in Proton-Proton-Kollisionen erzeugten Ereignisse sind voneinander unabhängige, statistische Pro-

zesse. Ihre Analyse ist daher besonders gut für einen verteilten Ansatz geeignet, da es im Gegensatz zu vielen anderen Anwendungen mit hohem Rechenbedarf kaum Kommunikationsbedarf zwischen den einzelnen Rechenschritten gibt und sich jedes Kollisionsereignis individuell betrachten lässt. Daher sind die benötigten Rechenschritte während der Rekonstruktion und Simulation trivial parallelisierbar, indem im einfachsten Fall jeweils ein komplettes Ereignis auf einem einzelnen Rechenkern bearbeitet wird (**Infokasten** „Simulation und Datenauswertung").

Seit 2001 ist das weltweite LHC Computing Grid Konsortium (WLCG) für die Entwicklung, den Aufbau und den Betrieb einer Computing-Infrastruktur für die Speicherung und Analyse der Daten der vier großen Experimente am LHC zuständig. Das Konsortium koordiniert alle teilnehmenden Rechenzentren sowie die einzusetzende Zugangs-, Verwaltungs- und Netzwerksoftware, die „Grid Middleware", welche die Nutzerschnittstelle zu den Ressourcen des Grid darstellt [4]. Sie stellt dabei die zentralen Komponenten bereit, die dafür sorgen, die Rechenanforderungen zu verteilen und die Grid-Ressourcen und Dienste zu überwachen und hinsichtlich der Leistung zu optimieren.[1)]

Als Organisationsform im Grid dienen „virtuelle Organisationen" (VO), die für ihre jeweiligen Nutzergruppen mit den Ressourcen-Betreibern Nutzungs- und Zugangsbedingungen aushandeln. Dazu gehören sowohl Rechen- oder Speicherressourcen, Servicequalität, Zuverlässigkeit und Sicherheit der angebotenen Dienste als auch eventuelle spezifische Arrangements hinsichtlich der Installation von Software oder des Betriebs von besonderen Diensten.

Eine hierarchische Struktur

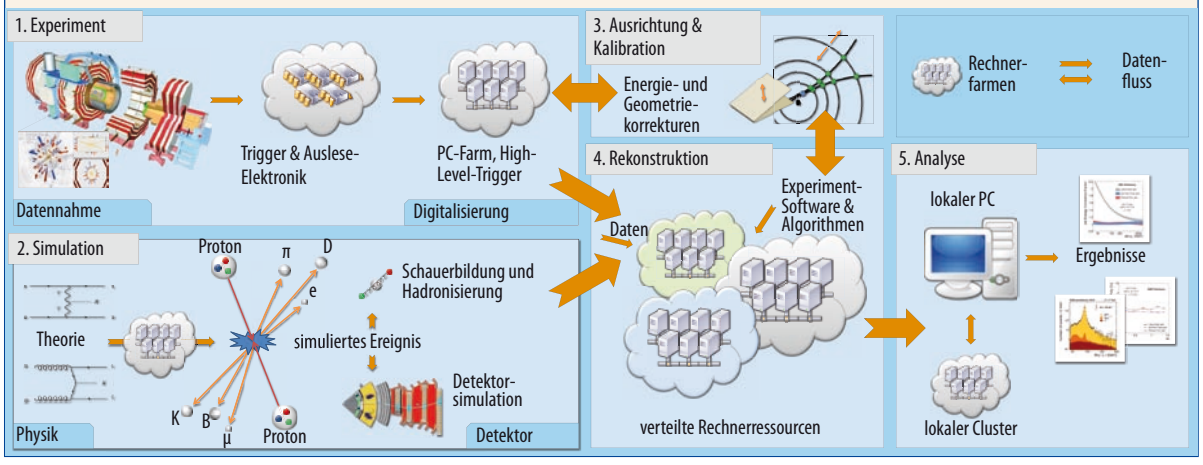
Basierend auf Studien aus den Jahren 1998 bis 2000 ist die Datenverarbeitung für die LHC-Experimente hierarchisch strukturiert und besteht aus verschiedenen Ebenen (engl. „Tiers") mit jeweils unterschiedlichen Aufgaben [5]. Die Speicherung der Experimentdaten und eine erste Rekonstruktion der von den Experimenten selektierten Ereignisse findet am Tier-0-Zentrum am CERN statt. Eine Kopie der Rohdaten sowie

1) Dabei sind große Grid-Initiativen wie EGEE („Enabling Grid for E-SciencE") in Europa oder „Open Science Grid" in den USA ebenso eingebunden wie die kleineren Ansätze „NorduGrid" in den nordeuropäischen Ländern oder das von der Alice-Kollaboration entwickelte „AliEn" (Alice Environment).

SIMULATION UND DATENAUSWERTUNG

Das Computing in der Teilchenphysik lässt sich in fünf typische Arbeitsschritte einteilen:

1. Die Datennahme am ***Experiment***, die durch geeignete nachgeschaltete Trigger-Stufen den enormen Rohdatenfluss auf verarbeitbare Datenmengen reduziert.
2. ***Simulationen*** von Kollisions-Ereignissen sind auch begleitend zur Datennahme notwendig, um die Signaturen möglicher neuer physikalische Theorien in den Detektoren zu beschreiben und aus den Daten nicht direkt messbare Untergrundbeiträge zu berechnen. Fortschritte in der Genauigkeit der Berechnungen machen es notwendig, diesen Schritt gegebenenfalls mehrfach zu wiederholen.
3. Zeitabhängige ***Kalibration***sfaktoren, die sich aus der Drift der Sensitivität von Detektorkomponenten oder aus sich ändernden Betriebsbedingungen des Beschleunigers ergeben, müssen bestimmt und in die eigentliche Rekonstruktion der Rohdaten zurückgefüttert werden. Diese Aufgabe ist sehr zeitkritisch und wird normalerweise auf Computing-Ressourcen direkt am Experiment ausgeführt.
4. Bei der ***Rekonstruktion*** entstehen zunächst aus den Rohdaten, also aus den in den verschiedenen Detektorzellen registrierten Signalen, die physikalisch relevanten Größen wie Teilchenspuren mit Impuls und Ladung oder Cluster aus Energiedepositionen in den Kalorimeterzellen. Im nächsten Schritt werden dann physikalische Objekte, d. h. identifizierte Teilchen oder Teilchenjets, gebildet. Direkt anschließend an diese Rekonstruktion findet auch eine Vorklassifizierung und Selektion der Ereignisse in verschiedene Datenströme statt, die der Ausgangspunkt der nachfolgenden Analysen sind. Mit fortschreitendem Verständnis der Detektoren und im Laufe der Zeit verbesserten Rekonstruktions- und Selektionsalgorithmen wird dieser Schritt mehrfach wiederholt. Dabei müssen auch die simulierten Daten mit den verbesserten Algorithmen neu rekonstruiert werden.
5. An letzter Stufe steht die statistische ***Analyse*** der Daten, die auf komprimierten Datensätzen beruht, in denen die jeweils für eine gegebene Analyse interessanten Ereignisse angereichert und mit im Vergleich zu Rohdaten-Ereignissen reduzierter Informationsmenge gespeichert sind. Optimierte Trennung von Signal und Untergründen, Vergleich mit theoretischen Modellen, Extraktion von physikalisch relevanten Messgrößen und die Aufbereitung der Daten in Form von Grafiken finden auf dieser Stufe statt.



die aus dem ersten Rekonstruktionsdurchlauf am CERN gewonnenen Daten werden an regionale Tier-1-Zentren weltweit verteilt, die für eine wiederholte Rekonstruktion der Rohdaten mit jeweils verbesserter Detektor-Kalibration, Rekonstruktionssoftware oder mit anderen Algorithmen verantwortlich sind. Diese sorgen außerdem für die Verteilung der rekonstruierten Ereignisse an zahlreiche nachgelagerte nationale Tier-2-Zentren. Die weitere Analyse und die Durchführung von Ereignissimulationen sind die primären Aufgaben dieser Tier-2-Zentren, die mit Instituts-rechnern und lokalen Arbeitsplatzrechnern verbunden sind. Die in vielfältigen Variationen verfügbaren Rechenanlagen an den Instituten selbst werden häufig als „Tier-3" bezeichnet, haben aber im LHC Computing Grid keine klar definierten Aufgaben und stellen auch nicht notwendigerweise Dienste oder Ressourcen nach außen bereit. Meist sind solche Zentren speziell auf die Bedürfnisse der Endphase von Datenanalysen optimiert, werden aber auch für Simulationsaufgaben eingesetzt und leisten so einen wichtigen Beitrag für die jeweilige experimentelle Kollaboration.

Derzeit umfasst das LHC-Grid zusätzlich zum Tier-0 am CERN elf Tier-1- und etwa 140 Tier-2-Standorte (**Abb. 2**). Die gesamte Rechenkapazität entspricht dabei 150 000 CPU-Kernen, von denen die Tier-2-Zentren etwa die Hälfte bereitstellen.[2)] Hinzu kommt eine Plattenspeicherkapazität von insgesamt etwa 100 Petabyte sowie weitere knapp 100 Petabyte an Bandspeicher (**Abb. 3**). Die Ressouren dieser weltweit größten Grid-Struktur werden im Jahr 2011 noch einmal um über 30 Prozent anwachsen. Die Gremien des WLCG stimmen jährlich die kommenden Ressourcen-Anforderungen der Experimente mit den jeweiligen Anbietern und Zuwendungsgebern ab.

Die Struktur in Deutschland

Institute der Teilchenphysik haben sich auf allen Ebenen der Entwicklung von Grid-Komponenten wesentlich beteiligt. Oft wurden sie dabei durch nationale Initiativen unterstützt, in Deutschland z. B. die D-Grid-Initiative des BMBF, die mittlerweile in die gemeinnützige D-Grid GmbH überführt wurde. Beim Aufbau einer leistungsfähigen Grid-Infrastruktur in Deutschland haben sich insbesondere die Helmholtz-Gemeinschaft deutscher Forschungszentren, das BMBF, die Max-Planck-Gesellschaft und einige Universitäten engagiert. Für die Einbindung der Kompetenzen an den Universitäten und deren Beteiligung am LHC-Grid war die Helmholtz-Allianz „Physik an der Teraskala" von großer Bedeutung, die neben Personal für Entwicklungsprojekte von Grid-Komponenten auch Hardware-Beschaffungen für das LHC-Grid an Universitäten finanzierte. Im Rahmen der Verbundforschung hat das BMBF Personal für experimentspezifische Aufgaben an den deutschen Zentren des LHC-Grids finanziert.

Das Karlsruher Institut für Technologie (KIT) betreibt mit GridKa[3)] das regionale Tier-1-Zentrum für Mitteleuropa, während am DESY in Hamburg und Zeuthen, bei der GSI in Darmstadt, der Max-Planck-Gesellschaft in München und an den Universitäten Aachen, Freiburg, Göttingen, München und Wuppertal Tier-2-Zentren angesiedelt sind. Viele der deutschen Universitäten nutzen PC-Cluster als Tier-3-Zentren, die teilweise über eine Anbindung ans Grid verfügen, ohne jedoch gegenüber WLCG Ressourcen fest zuzusagen. Insbesondere die D-Grid-Initiative hat zum Aufbau solcher nationaler Grid-Ressourcen beigetragen. Auf den Instituts-Clustern findet im Allgemeinen die tägliche Arbeit der Wissenschaftler statt, die dort interaktiv oder zumindest mit kurzer Antwortzeit auf Daten zugreifen können, die zuvor im Grid prozessiert und selektiert wurden.

DESY betreibt darüber hinaus besonders für die Endphase von Datenanalysen die „National Analysis Facility" (NAF).[4)] Die NAF bietet ein verteiltes Filesystem (AFS) sowie Speicherplatz auf einem schnellen, parallelen Filesystem (Lustre). Von den Rechnern der NAF aus ist der volle Tier-2-Datenbestand des DESY sichtbar. Der deutsche Beitrag zu den gesamten Grid-Ressourcen für den LHC beträgt etwa 15 Prozent im Tier-1- bzw. knapp 10 Prozent im Tier-2-Bereich.

Das Grid in der Praxis

Aus Sicht des Computing verlief das erste Jahr der Datennahme und Analyse am LHC äußerst erfolgreich. Die beiden großen Experimente ATLAS und CMS haben wöchentlich mehrere Millionen Grid-Jobs bearbeitet (**Abb. 4**). Die aufgebaute Infrastruktur hat sich bewährt und die vorgesehenen Aufgaben für die Datenprozessierung, Datenselektion und Verteilung in hervorragender Weise bewältigt.

Zuverlässigkeit

In den Jahren vor dem LHC-Start war die schrittweise aufgebaute Grid-Struktur in immer komplexeren Probeläufen getestet worden. Automatisierte Überwachungsmechanismen sorgen bei einer Fehlfunktion für ein zeitnahes Feedback an die Tier-1-Standorte, die kritische Dienste rund um die Uhr unterstützen. Für

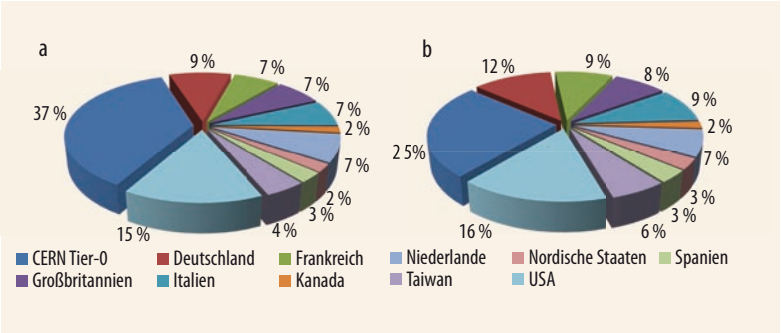


Abb. 3 Überblick über die geografische Verteilung der CPU- (a) bzw. Plattenplatz-Ressourcen (b) für Tier-0 und Tier-1 im Jahr 2010. Noch einmal der gleiche Beitrag an Rechenleistung wird von den Tier-2-Zentren bereit gestellt (nicht gezeigt). Der deutsche Anteil daran ist 8 %.

2) Zur Messung der Leistungsfähigkeit eines Rechners für typische Anwendungen aus der Teilchenphysik wird die Einheit HEP-SPEC06 verwendet; ein aktueller Kern einer CPU mit 3 GHz Taktrate entspricht etwa 10 HEP-SPEC06.

3) www.gridka.de

4) naf.desy.de

Tier-2-Zentren beschränkt sich diese Unterstützung allerdings auf die normale Arbeitszeit. Im ersten Jahr des LHC-Betriebs lag im Tier-1-Bereich der durch das Versagen von Grid-Komponenten verursachte Anteil an Abbrüchen von Rechenanforderungen oder Datentransfers bei wenigen Prozent. Im Durchschnitt aller Tier-2-Zentren liegt dieser Wert allerdings höher und schwankt sehr stark zwischen den besten Zentren und denen am Ende der Skala. Die Zuverlässigkeit der Grid-Infrastruktur wird in Zukunft sicher weiter steigen.

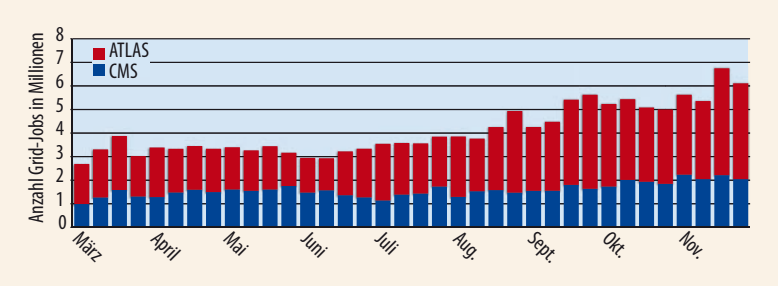


Abb. 4 Anzahl der Grid-Jobs, die bei den zwei größten LHC-Experimenten ATLAS und CMS im Jahr 2010 pro Woche angefallen sind.

Datenverteilung an die Tier-1-Zentren

Die vom CERN an die Tier-1-Zentren exportierten Datenmengen haben im ersten Jahr des LHC-Betriebs die ursprünglichen Planungen sogar noch übertroffen. So umfasste das durch das CMS-Experiment transferierte Datenvolumen im Jahr 2010 knapp 3 PByte. Zu den Rohdaten des Experiments kommt noch ein erheblicher Anteil an simulierten Kollisions-Ereignissen, die in Tier-2-Zentren erzeugt und ebenfalls an die Tier-1-Standorte zur Langzeitspeicherung geschickt wurden. Bei CMS waren dies weitere etwa 3 PByte. Die ATLAS-Kollaboration hat im selben Zeitraum insgesamt etwa 8 PByte an den Tier-1-Zentren gespeichert. Allein auf den Plattensystemen am deutschen Zentrum GridKa liegen derzeit über 5 PByte an Daten der LHC-Experimente.

Die Rechenleistung der Tier-1-Zentren wurde hauptsächlich für die wiederholte Rekonstruktion der Daten genutzt. Bedingt durch das relativ kleine Datenvolumen in der Anfangsphase waren sogar deutlich häufiger Rekonstruktionsdurchläufe möglich als ursprünglich vorgesehen bzw. als sie in Zukunft möglich sein werden. Dadurch konnten die Ergebnisse der ersten Datenanalysen in die darauffolgenden Datenrekonstruktionen einfließen, was sehr wesentlich zu den qualitativ hochwertigen und bereits veröffentlichten Analysen der LHC-Experimente beitrug.

Nutzung der Tier-2-Zentren

Die Hauptlast der Datenanalyse durch die Physiker tragen die Tier-2-Zentren. Neben der Produktion von simulierten Datensätzen sehen die Computing-Modelle von ATLAS und CMS vor, dass ein Tier-2-Zentrum vorselektierte und im Umfang reduzierte Datensätze für die Datenanalyse bereithält, auf die Nutzer mit Grid-Werkzeugen zugreifen. Im Jahr 2010 wurde routinemäßig ein Durchsatz von etwa 100 000 solcher Analyse-Jobs pro Tag für jedes der großen Experimente CMS und ATLAS erreicht (**Abb. 5**).

Sobald eine neue Rekonstruktion der Daten an den Tier-1-Zentren durchgeführt wird, müssen die an den Tier-2-Zentren gespeicherten Daten ersetzt werden. Sie sind zu diesem Zweck in einem privaten WAN-Netzwerk mit einer Bandbreite von 10 GBit/s vernetzt und erhalten so die neuen Daten zeitnah über diese Verbindungen. Bei der Verteilung an die Tier-2-Zentren verfolgt ATLAS dabei ein regionales Modell, bei dem jedes Tier-2 einem Tier-1 zugeordnet ist, von dem es seine Datensätze bezieht. Bei CMS geschieht diese Verteilung weltweit, d. h. jedes Tier-2 kann auf die Datensätze an jedem CMS-Tier-1 zugreifen; es sind dabei überdies auch Transfers von einem Tier-2 an ein anders möglich. Auch bei ATLAS werden sich die Mechanismen zur Datenverteilung in Zukunft in diese Richtung entwickeln. Eine solche Strategie garantiert die schnellstmögliche Verteilung der Daten an alle Tier-2-Zentren, allerdings entsteht so eine hohe Netzwerklast, insbesondere auch im transatlantischen Verkehr. Die Netzwerkbandbreiten müssen also in der nahen Zukunft aufgrund der steigenden Anforderungen angepasst werden.

Endphase der Datenanalyse

Mittlerweile ist die Akzeptanz von Grid-basierten Analysen bei den Nutzern deutlich gestiegen, insbesondere auch durch die Entwicklung von automatisierten und teilweise mit grafischer Oberfläche versehenen Grid-Schnittstellen, die das Abschicken einer großen Menge von Rechenanforderungen und das Abholen der Ergebnisse erheblich erleichtern, da sie z. B. Datenbanken der Experimente nutzen, um vorhandene Datensätze automatisch aufzufinden (**Infokasten** „Ablauf einer Rechenanforderung“).

Der digitale Grid-Ausweis

In einem zweistufigen, aber automatisierten und zuverlässigen Registrierungsprozess erhält jeder an

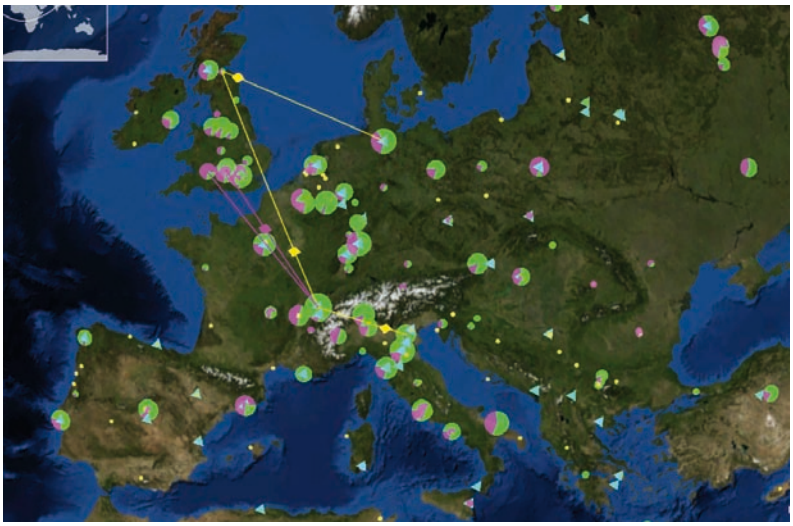


Abb. 5 Die europäischen Zentren im LHC-Grid. Die Verbindungslinien zeigen gerade aktive Transfers von Rechenanforderungen.[5)]

5) Grid Real Time Monitor, http://rtm.hep.ph.ic.ac.uk/

einem LHC-Experiment beteiligte Physiker Zugriff auf die weltweit verteilten Ressourcen des LHC-Grids. Im ersten Schritt bestätigt eine regionale Zertifizierungsstelle die Identität des Grid-Nutzers und stellt ihm einen eigenen Zertifizierungsschlüssel bereit, mit dem dieser im zweiten Schritt bei einer virtuellen Organisation – in diesem Fall bei seinem LHC-Experiment – registriert wird und so dessen Grid-Ressourcen nutzen kann und Zugriff auf die jeweiligen Daten erhält.

Arbeiten im Grid

Für junge Physiker in den LHC-Kollaborationen ist es selbstverständlich geworden, die weltweit verteilten Ressourcen des Grid zu nutzen. Sie erhalten Zugriff auf den vollen Datensatz ihres jeweiligen Experiments und können darüber hinaus auch die im Hinblick auf bestimmte interessante Signaturen angereicherten Datensätze ihrer Arbeitsgruppen nutzen. Dies erlaubt sowohl kollaboratives Arbeiten als auch die möglichst freie Umsetzung eigener Analyse-Ideen.

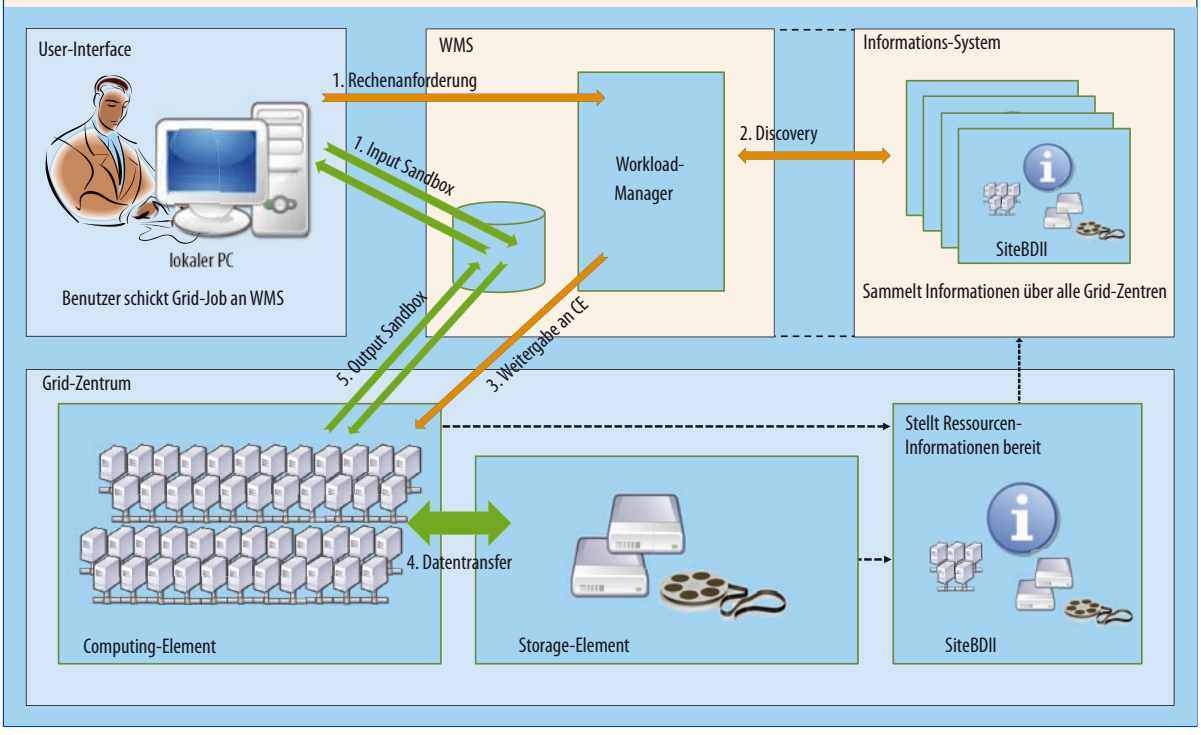
Für die Endphase der Datenanalyse werden typischerweise hochangereicherte und auf die jeweils wesentlichen Informationen beschränkten Datensätze aus den vorselektierten Daten an den Tier-2-Zentren extrahiert und zur weiteren Analyse an das jeweilige Heimatinstitut transferiert. Dort finden Optimierungen der Datenselektion, Visualisierung und statistische Analyse der Daten statt. Diese Schritte müssen sehr häufig wiederholt werden, wobei die Dauer eines solchen Laufs über die Daten im kompetitiven internationalen Umfeld von entscheidender Bedeutung für den Erfolg einer Analyse ist. Aus Sicht des Computing sind daher an dieser Stelle deutlich höhere Schreib- und Lese-Raten nötig als bei der Datenrekonstruktion. Viele Zentren setzen deshalb für diese Anwendungen schnelle, parallele Filesysteme ein.

Was bringt die nahe Zukunft?

Im Regelfall sind zwischen den Datennahmephasen am LHC längere Pausen für Wartungsfenster und Hardware-Upgrades vorgesehen. Mit der Entscheidung, den LHC ohne längere Unterbrechung bis Ende 2012 zu betreiben, steht dem LHC-Grid eine weitere Bewährungsprobe bevor. Der bisher oft noch sehr personalintensive Betrieb muss optimiert werden, ohne dass es zu Einbußen bei der Leistungsfähigkeit kommt.

ABLAUF EINER RECHENANFORDERUNG AUF DEM GRID

1. Jeder Nutzer benötigt zunächst Zugang zu einem „User Interface“ (UI), einer Kommandozeilen-basierten Anwendung für die Kommunikation mit den Komponenten des Grids. Über das UI werden alle benötigten Dateien in einer „Input Sandbox“ sowie eine Beschreibung der Rechenanforderung selbst in einer speziellen Beschreibungssprache („Job Description Language“, JDL) an den „Workload Management Service“ (WMS) übergeben, der für die Lastverteilung und Ressourcen-Lokalisierung im Grid verantwortlich ist.
2. Auffinden eines Grid-Standorts, der die benötigten Ressourcen bereitstellen kann – d. h. insbesondere freie Rechenknoten und benötigte Eingabedaten. Jeder Grid-Standort veröffentlicht diese Informationen über den „SiteBDII“ Service.
3. Zuweisung an eine geeignete Rechnerfarm im Grid („Computing-Element“, CE) und Ausführung der Rechenanforderung auf einem an das CE angeschlossenen Rechnerknoten.
4. Während der Abarbeitung stehen den Rechenanforderungen Speicherelemente („Storage-Element“, SE) – entweder lokal oder über Netzwerk – zur Verfügung, um größere Mengen an Eingabedaten zu beziehen oder Ausgabedaten zu speichern.
5. Das Ergebnis wird in der „Output Sandbox“ zum WMS kopiert, wo es der Nutzer abholen kann.



Das zu verarbeitende Datenvolumen wird schneller anwachsen als die durch technologischen Fortschritt bei gleichen Kosten zu realisierende Vergrößerung der Speicherressourcen. Daher wird es nicht mehr möglich sein, Daten sozusagen „auf Vorrat" an alle Zentren zu verteilen, sondern neue Mechanismen zur Datenverteilung sind gefragt, um nur benötigte und tatsächlich genutzte Datensätze bei Bedarf zu transferieren. Selten genutzte Datensätze könnten dann eventuell auch über Netzwerk prozessiert werden, ohne eine lokale Kopie vorzuhalten. Dies würde zu deutlich erhöhten Anforderungen an die Netzwerk-Infrastruktur führen, die derzeit allerdings (noch) kein limitierender Faktor ist.

Mit dem Beginn dieses Ressourcen-limitierten Betriebs der LHC-Infrastruktur wird das optimierte Nutzen der vorhandenen Ressourcen mehr und mehr in den Fokus rücken. Dazu gehört es, Fehlerraten und damit die Notwendigkeit von Wiederholungen der entsprechenden Rechenschritte zu reduzieren und wenig genutzte Ressourcen zu identifizieren und verstärkt einzubinden. Die Datenanalyse an den Heimatinstituten war häufig noch geprägt von einer klassischen Herangehensweise, bei der auf dem Grid erzeugte Datensätze letztlich auf institutseigenen Ressourcen analysiert wurden. Die Abhängigkeit der Grid-Middleware von einem dedizierten Betriebssystem und deren komplexe Struktur hat es Instituten nicht leicht gemacht, ihre eigenen Ressourcen auf einfache Weise in das Grid einzubinden. Einen Ausweg bietet die auf immer breiterer Basis verfügbare Technik der Virtualisierung, die es erlaubt, eine Standard-Hardware zu emulieren sowie darauf eine nahezu beliebige Betriebssystemumgebung bereit zu stellen. Anbieter von Ressourcen sind damit von der Notwendigkeit befreit, zur Unterstützung der Teilchenphysiker das von WLCG vorgegebene Betriebssystem auf ihrer Hardware zu installieren.

Eine zunächst kommerziell ausgerichtete Variante der verteilten Nutzung von Rechnerressourcen trägt den Namen „Cloud Computing" und basiert auf Web-Diensten und der Abstraktion von der eigentlichen Hardware- und Betriebssystemumgebung durch Virtualisierungstechniken, die insbesondere Sicherheitsaspekte, dynamische Skalierbarkeit, Ressourcenverwaltung und Abrechnung des Verbrauchs berücksichtigen. Es existieren offene Implementierungen der gängigsten Cloud-Schnittstellen, mit denen der kostengünstige Aufbau „privater Clouds" möglich ist.

Der Weg für die Teilchenphysik scheint bereits vorgezeichnet: Die Zukunft wird ein Verschmelzen der Grid-Infrastruktur mit Komponenten des Cloud-Computing bringen. Anwendungen mit hohem Rechenbedarf, aber geringen Anforderungen an die Datenrate sind schon jetzt für die Ausführung auf virtuellen, dynamisch zugewiesenen Ressourcen geeignet. Die Rekonstruktion großer Datenmengen oder datenintensive Analysen werden zunächst spezialisierten Zentren mit Bandspeicher oder besonders leistungsfähigen Plattensystemen vorbehalten bleiben. Wenn sich die Kosten für die Datenübertragung im Netz weiter verringern, ist langfristig aber auch die Trennung von Datenspeichern und Rechenleistung vorstellbar. Erste Überlegungen zu einer weltweiten Dynamisierung des Datenmanagements auf dem existierenden Grid der Tier-1- und Tier-2-Zentren weisen schon in diese Richtung.

Die Computing-Gruppen aller vier Experimente und auch das WLCG-Team und die beteiligten Grid-Standorte bereiten sich auf die bevorstehende Herausforderung vor. Die sehr positiven Erfahrungen des letzten Jahres lassen erwarten, dass das LHC-Grid den Anforderungen der kommenden Datenanalysen gewachsen sein wird. Dies ist eine wichtige Voraussetzung für das Auffinden neuer und aufregender Physik am LHC, auf die alle Beteiligten hoffen.

Literatur

[1] International Conference on High Energy Physics – ICHEP 2010, http://www.ichep2010.fr/

[2] *V. Lindenstruth*, Physik Journal, Januar 2011, S. 23

[3] *I. Foster* und *C. Kesselman*, The Grid: Blueprint for a new Computing Infrastructure, Morgan Kaufmann Publishers Inc., San Francisco (1998)

[4] LHC Computing Grid – Technical Design Report, Version: 1.04, 2005, LCG-TDR-001, CERN-LHCC-2005-024, http://lcg.web.cern.ch/

[5] *M. Aderholz* et al., MONARC: Models of Networked Analysis at Regional Centres for LHC Experiments, 2000, CERN-LCB-2000-001, KEK-2000-8

DIE AUTOREN

Günter Quast (FV Teilchenphysik) hat in Siegen Physik studiert und 1988 über CP-Verletzung beim Zerfall von K-Mesonen promoviert. Anschließend forschte er am CERN sowie an der Universität Mainz, wo er sich 1998 habilitierte. Seit 2001 ist er Professor am Karlsruher Institut für Technologie (KIT, ehemals Universität Karlsruhe). Quast ist Mitglied der CMS-Kollaboration, seine Forschungsschwerpunkte liegen neben dem Grid-Computing auf elektroschwacher Physik sowie der Suche nach dem Higgs-Boson.

Armin Scheurer (FV Teilchenphysik) hat in Karlsruhe Physik studiert und 2008 über die Identifikation von B-Quark-Jets bei CMS promoviert. Seither ist er Postdoc am Karlsruher Institut für Technologie mit Schwerpunkt Computing, leitet die Gruppe zur Betreuung von experiment-spezifischen Diensten für das CMS-Experiment am GridKa und koordiniert Grid- und Cloud-Computing-Projekte am Institut für Experimentelle Kernphysik.

