

Zusammenfassung der Strategiediskussion zum Computing in der HL-LHC-Ära

Präambel

Am 6. und 7. Mai 2020 wurden strategische Fragen zum Computing in Deutschland für die kommenden 10 Jahre auf einem offenen Workshop von Wissenschaftlern aus den Gebieten Teilchen-, Hadronen- und Kern-, Astroteilchenphysik und Astronomie sowie den beteiligten Rechenzentren diskutiert. Die Agenda mit den Vorträgen und Fragen ist auf der Indico-Seite <https://indico.physik.uni-muenchen.de/event/33/> zu finden. Dieses Dokument fasst die Ergebnisse der Diskussion zusammen.

Durch stetig weiterentwickelte Experimente für die Erforschung von Universum und Materie (ErUM) werden immer größere Forschungsdatensätze bis in den Exabyte-Bereich gewonnen, die nur mit entsprechenden Computing-Systemen aufgezeichnet und ausgewertet werden können. Eine detaillierte Darstellung der Herausforderungen ist unter anderem im Strategiepapier der ErUM-Communities "[Challenges and Opportunities of Digital Transformation in Fundamental Research on Universe and Matter](#)" zu finden. Die Notwendigkeit der dort empfohlenen Maßnahmen, insbesondere bezüglich des Einhergehens von Hardware-Erweiterungen mit besseren Software-Tools und Analyse-Methoden wurde auf dem Workshop bekräftigt und in einigen Punkten konkretisiert. Dabei lag der Fokus auf dem Computing für den HL-LHC und es wurden nicht alle Aspekte des Themas abgedeckt.

Herausforderungen

In wenigen Jahren wird die Zunahme an Daten und Rechenaufwand deutlich das übersteigen, was bei konstantem jährlichen Budget (Flat-Budget) durch technologischen Fortschritt bei Speichermedien und CPUs zur Verfügung gestellt werden kann. Für die High-Luminosity-Phase des LHC gehen die aktuellen Extrapolationen von ATLAS und CMS von einem Faktor 4-10 gegenüber dem Flat-Budget aus. Der Faktor hängt von erreichbaren Optimierungen der Software, der Computing-Modelle und der Ressourcennutzung ab. Niedrige Faktoren, gegebenenfalls sogar kleiner als 4, können nur durch erheblich gesteigerten Forschungsaufwand erreicht werden. Auch der Ressourcenbedarf von FAIR, Belle II, der Astroteilchenphysik, der Astronomie und der Theorie wird in den nächsten Jahren signifikant ansteigen. Zum Beispiel wird der Bedarf an Computing-Ressourcen für die Leuchtturm-Experimente der Astroteilchenphysik in Deutschland bis zum Jahre 2028 um circa einen Faktor 10 zunehmen.

Mit dem Betriebsbeginn des HL-LHC 2027 wird es in jedem Fall einen sprunghaften Anstieg des Ressourcenbedarfs gegenüber Run 3 geben. Um den Anstieg bewältigen zu können, muss spätestens 2024 eine verlässliche Bedarfsabschätzung vorliegen und mit dem Aufbau der Ressour-

cen an Zentren begonnen werden. Die Voraussetzung hierfür ist, dass schon jetzt die Grundlagen für das Computing in der HL-LHC-Phase gelegt werden und gleichzeitig der Betrieb der Infrastruktur für die laufenden Experimente sichergestellt wird.

Aktuelle Situation

Durch Optimierungen der Software und der Computing-Modelle konnte der Faktor zum Flat-Budget im Vergleich zu ursprünglichen Schätzungen bereits deutlich reduziert werden. Reduzierungen um ein bis zwei Größenordnungen im Ressourcenbedarf haben auch ALICE und LHCb mit signifikanter deutscher Beteiligung nach ihren Upgrades erzielt, so dass deren Bedarf für Run 3 nur etwa 20-50% oberhalb des Flat-Budgets liegt. Diese Zahlen zeigen, dass der eingeschlagene Weg von frühzeitigen Investitionen in Forschung direkt zu einer nachhaltigen Reduktion von Investitionskosten über die gesamte Laufzeit des LHC führt. Hiervon profitieren auch die existierenden und geplanten Experimente in den angrenzenden Forschungsfeldern.

In internationalen Kollaborationen wird erwartet, dass Deutschland einen Anteil der benötigten Ressourcen bereitstellt, der dem Anteil beteiligter Wissenschaftler entspricht. Die Finanzierung der benötigten Ressourcen in Deutschland für die LHC-Experimente und Belle II ist momentan an den Helmholtz-Zentren und an den Universitäten bis 2021 gesichert.

Strategie

Die oben benannten Herausforderungen machen erhebliche Anstrengungen zur effizienten Nutzung von Computing-Ressourcen notwendig, um insbesondere das HL-LHC-Physikprogramm durchführen zu können. Die Teilnehmer des Workshops unterstützen nachdrücklich eine starke deutsche Beteiligung an den folgenden Themen:

- (Weiter)Entwicklung und Optimierung der Experimentsoftware, inklusive der Nutzung neuer Architekturen wie GPUs
- (Weiter)Entwicklung und Optimierung gemeinsamer Softwarepakete
- (Weiter)Entwicklung und Optimierung von Komponenten zur effizienten Nutzung von (opportunistischen) heterogenen Ressourcen bei reduziertem Aufwand für Installation und Betrieb, was auch die Nutzung kommerzieller Ressourcen bei Peak-Last beinhaltet.

Die geplanten Entwicklungsarbeiten knüpfen an die Vorarbeiten aus dem ErUM-Data-Pilotprojekt an und benötigen erhebliche zusätzliche Finanzmittel. Dabei werden Community-übergreifende Verbände zum Themenkomplex "enabling Exabyte-Computing" anvisiert. Sinnvoll erscheint eine thematische Trennung in Computing & Infrastruktur und wissenschaftliche Software & Analyse.

Trotz Optimierungen werden signifikante Hardware-Ressourcen benötigt, die spätestens ab 2024 aufgebaut werden müssen. Die Helmholtz-Zentren KIT, DESY, und GSI haben auf dem Workshop deutlich gemacht und werden dabei von den Teilnehmern des Workshops unterstützt, dass sie es als ihre Aufgabe sehen, einen großen Teil der benötigten Ressourcen als deutschen Beitrag zu WLCG bereitzustellen, und dafür entsprechende Investitionsmittel benöti-

gen. Eine breite und synergetische Nutzung der Ressourcen unter Einbeziehung der Bedarfe der angrenzenden Communities ist dabei Teil der Strategie. Universitäre Computing-Ressourcen, insbesondere wenn sie als Teil eines größeren Rechenzentrums betrieben werden, stellen weiterhin einen wichtigen Beitrag dar und die dort betriebene Ausbildung dient als Multiplikator von Computing-Expertise. Die Einbindung opportunistischer Ressourcen ist durch förderierte Infrastrukturen und Technologien zu ermöglichen. Grundvoraussetzung ist eine leistungsstarke Netzwerkinfrastruktur.

Bei der Computing-Infrastruktur geht der Trend klar zu einem Data-Lake-Szenario, bei dem große (förderierte) Zentren die Daten speichern und kleinere Zentren auf diese über Netzwerk und Zwischenspeicher zugreifen. Diese Entwicklungen werden von der deutschen Community mitgetragen und maßgeblich mitentwickelt. GridKa, DESY und GSI haben national und international eine herausragende Rolle und werden als unverzichtbar für den Data-Lake-Backbone und die Datenarchivierung angesehen. Das Wissen über experimentenspezifische Komponenten und Workflows liegt oft bei den Universitäten, die in diesem Bereich auch Verantwortlichkeiten übernehmen und somit für das Funktionieren der Infrastruktur ebenfalls wesentlich sind.

Die große Mehrheit der Workshopteilnehmer sieht die Notwendigkeit, neben lokalen Analyse-Clustern (insbesondere an Universitäten) nationale und internationale Analyse-Facilities (DESY und GSI) mit aufeinander optimiertem Design von Hardware und Software in Deutschland einzurichten; die entsprechenden Pläne dazu werden unterstützt. Dies ist ein wichtiger Schritt hin zu einer deutschen Science-Cloud. Für die Realisierung dieses Konzepts sind jedoch weitere Entwicklungsarbeiten erforderlich, die auf Vorarbeiten und Erfahrungen der deutschen Community aufbauen.

Neue Technologien wie maschinelles Lernen und künstliche Intelligenz spielen eine wichtige Rolle bei der Reduktion des Ressourcenbedarfs und der Maximierung des wissenschaftlichen Erkenntnisgewinns aus aufgezeichneten Daten. Die Möglichkeiten, die Quantencomputing bieten wird, sind noch nicht abzuschätzen. Auch wenn derzeit davon auszugehen ist, dass Quantencomputing keine große Rolle für den HL-LHC spielen wird, sollte es Expertise in der Community geben, um das Potenzial der neuen Technologie beurteilen zu können.

Die Finanzierung von Entwicklungsarbeiten und Hardware für das "Exabyte Computing" und experimentübergreifende Ansätze sollte in der Förderlinie ErUM-Data erfolgen. Allerdings gibt es auch Themen, die über DFG-Anträge, oder innerhalb von Forschungsorganisationen (Helmholtz Gemeinschaft und Max-Planck-Gesellschaft) finanziert werden könnten. Für die Erfüllung der experimentenspezifischen Aufgaben im laufenden Betrieb der Experimente stellt ErUM-Pro den richtigen Rahmen dar.

Kooperation und Nachhaltigkeit

Wie auf dem Workshop deutlich wurde, stehen viele Experimente vor ähnlich großen Herausforderungen und auch wenn ein Teil der Arbeiten experimentenspezifisch sein muss, wird eine Zu-

sammenarbeit ausdrücklich begrüßt. Eine Finanzierung gemeinsamer Projekte in Hard- und Software von Teilchenphysik, Hadronen- und Kernphysik, Astroteilchenphysik und Astronomie, z.B: durch ErUM-Data, wird gewünscht. Das Pilotprojekt hat gezeigt, dass Experiment- und Community-übergreifende Lösungen erfolgreich gemeinsam entwickelt werden können und sollten. Eine Zusammenarbeit mit entsprechenden nationalen Aktivitäten wie der NFDI und internationalen Initiativen wie HSF, EOSC und ESCAPE ist dabei erforderlich. Die ErUM-Data-Dachorganisation kann eine wichtige Rolle bei der Koordinierung zwischen den ErUM-Communities übernehmen. Entwicklungen in den oben genannten Wissenschaftsbereichen werden längerfristig auch den anderen ErUM-Communities zugute kommen.

Die Bedarfe an Ressourcen sollen Community-übergreifend ermittelt und den Geldgebern präsentiert werden. Die Möglichkeiten von Zusammenarbeit mit Informatikern, Mathematikern und Industrie ist projektabhängig zu eruieren. Hier kann die ErUM-Data Dachorganisation helfen, Kontakte herzustellen. Wie beim Detektorbau ist es auch im Bereich Computing und Software unverzichtbar, dass ausreichend Expertise in der Community vorhanden ist.

Die Anerkennung von Leistungen im Bereich Forschungsinfrastrukturen und Forschungssoftware nimmt zu. Diese Entwicklung wird sehr begrüßt, da sie für erfolgreiche und nachhaltige Projekte sehr wichtig ist. Solche Projekte erfordern eine lange Laufzeit und ständige Wartung und Support, wofür es oft schwierig ist eine adäquate Finanzierung zu finden. Insbesondere die Kontinuität der Forschungstätigkeiten über Förderperioden hinweg ist ein wichtiger Aspekt, der zu einer stärkeren und weithin sichtbaren deutschen Beteiligung an internationalen Projekten und auch zu einer Erhöhung des Frauenanteils im Forschungsfeld beitragen könnte. Das im ErUM-Data-Strategiepapier vorgeschlagene Tenure-Track-Programm wird als geeignetes Mittel zur Stärkung der Nachhaltigkeit gesehen.