GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

# Schlussbericht

Verbund: 05P2021 - Run 3 von ALICE am LHC


| | |
|---|---|
| Zuwendungsempfänger: | Johann Wolfgang Goethe-Universität Frankfurt am Main |
| Projektleitung: | Prof. Dr. Volker Lindenstruth |
| E-Mail: | voli@compeng.de |



| | |
|---|---|
| Förderkennzeichen: | 05P21RFCA2 |
| Förderzeitraum: | 01.07.2021 - 30.06.2024 |
| Zuwendung: | 1.545.639,76 € |
| Projektträger: | Projektträger DESY |


Zusätzlicher Kontakt:
Zusätzlicher Name:

Genutzte Großgeräte:

| Labor | Gerät | Experiment |
|---|---|---|
| CERN | LHC | ALICE |

| | |
|---|---|
| Diplomarbeiten: | 0 |
| Dissertationen: | 1 |
| Habilitationen: | 0 |
| Referierte Publikationen: | 270 |
| Andere Veröffentlichungen: | 0 |
| Patente: | 0 |
| Bachelorarbeiten: | 1 |
| Masterarbeiten: | 3 |
| Staatsexamen: | 0 |

Dieser Bericht wurde beim Projektträger über einen individuellen Online-Zugang vom Projektleiter eingereicht und am 14.03.2025 09:27 für eine Veröffentlichung freigegeben.

# Schlussbericht

| | |
|---|---|
| Zuwendungsempfänger: | Goethe-Universität Frankfurt am Main |
| Projektleitung: | Prof. Dr. Volker Lindenstruth |
| Verbund: | 05P21RFCA2: Verbundprojekt 05P2021 (ErUM-FSP T01)- Run 3 von ALICE am LHC: Fertigstellung und Inbetriebnahme der EPN-Farm” |
| Thema: | Arbeitspakete – Prof. Dr. U. Kebschull<br>o Entwicklung schneller Datenvorverarbeitung mittels High-Level-Synthese im FLP mit Schwerpunkt auf die Time Projection Chamber (WP6, TPC Readout)<br><br>Arbeitspakete – Prof. Dr. V. Lindenstruth<br>o Aufbau und Betrieb der EPN Farm<br>o Einsatz beim Strahlbetrieb<br>o übergreifende Qualitätssicherung |



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

# Zusammenfassung

Die Arbeiten der Antragssteller wurden im Kontext der im ErUM-FSP T01 organisierten deutschen ALICE Universitätsgruppen in Frankfurt, Münster, Heidelberg, München, Bonn, Bielefeld und Tübingen sowie der GSI durchgeführt. Sie stellen einen erheblichen Beitrag zum ALICE Experiment dar. Hier sind insbesondere die TPC, der TRD und die EPN-Rechenfarm zu nennen. Die Projektleitung dieser drei Teile von ALICE liegt bei den deutschen Gruppen. Darüber hinaus gab es eine intensive Zusammenarbeit mit der FLP und der PDP Gruppe am CERN im Kontext des Rechnerbetriebes und besonders im Bereich der Entwicklung der schnellen und effizienten Ereignisrekonstruktionssoftware auf GPUs. Die EPN-Rechenfarm ist bezüglich des sehr konsequenten Einsatzes von GPU-Rechenbeschleunigern am CERN führend. Im Kontext der Validierung der verschiedenen Softwarekomponenten der ALICE Subdetektoren wurde mit den entsprechenden Verantwortlichen der Detektorgruppen, insbesondere der TPC, zusammengearbeitet und Hilfestellung bei der Softwareentwicklung geleistet. Es besteht eine Validierungsinfrastruktur die alle neue Software zu durchlaufen hat um sicherzustellen, dass die Software möglichst fehlerfrei ist.

Das ALICE Experiment am Large Hadron Collider des CERN wurde während des Run 3 betrieben und weiterentwickelt. Die ALICE Event Processing Farm (EPN) ist eine online Farm, die die Daten des Experiments in Echtzeit analysiert und komprimiert. Hierbei werden die Daten des Detektors mit einer Rate von knapp einem Terabyte pro Sekunde verarbeitet. Eine maximale Verarbeitungsrate von 1,3 Terabyte pro Sekunde konnten demonstriert werden. Hierbei wird eine volle Ereignisrekonstruktion durchgeführt. Ein wesentliches Element dieser on-line high-throughput HPC Farm ist der konsequente Einsatz von Graphikkarten als Hardware Beschleuniger um die Kosten zu reduzieren. Mittlerweile sind fast 100% der Software des synchronen Rechnens (online) auf Graphikkarten lauffähig. Insgesamt kann derselbe Quellcode sowohl auf verschiedenen GPUs aber auch auf CPUs laufen. Es hat sich gezeigt, dass die EPN Farm das Siebenfache kosten würde wenn auf GPUs verzichtet worden wäre. Die EPN Farm implementiert 2800 AMD MI50/MI100 Server GPUs, 24.640 physikalische AMD CPU Kerne, 200 TB Hauptspeicher und ein 100 Gb/s InfiniBand Netzwerk. Die Farm besteht aus 350 individuellen Servern. Die EPN Farm wird auch für das asynchrone (offline) Rechnen verwendet. Wesentliche Entwicklungen waren der TPC Readout, die Monitoring Umgebung, die 4D Spurrekonstruktion, die Qualitätskontrollinfrastruktur für die Datenaufnahme, die Entwicklung tragfähiger und effizienter Datenschemata. Ein zentraler Baustein der EPN Farm ist die Daten- und Lastverteilung, die die gesamten Datenströme in der EPN Farm steuert. Sie wurde während der Förderperiode zusammen mit aller notwendigen Software fertiggestellt und ist seitdem im Einsatz.



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

# Bericht

## Subprojekt Prof. Dr. Udo Kebschull:

### 1 Aufgabenstellung und Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Arbeitspaket WP6 ist eine thematische Fortführung zu den Arbeiten während der Umrüstphase zu Run3.

Die immer höher werdende Datenrate der Detektorauslese erfordert eine immer höher werdende Reduktion der ausgelesenen Daten in Echtzeit. Das heißt immer komplexere Algorithmen sind notwendig, die traditionell immer noch in VHDL implementiert werden. Mit zunehmender Komplexität wird das immer schwieriger und zeitaufwendig. Um den wachsenden Anforderungen gerecht zu werden, wurde deshalb an einer Methode gearbeitet mit deren Hilfe Datenvorverarbeitungsalgorithmen in Modern C++ und HLS entwickelt, getestet und als IP in bestehende FPGA-Ausleseferfirmware integriert werden kann.

Jedes der LHC-Experimente, so auch das ALICE, muss am „Karlsruher Institut of Technology (KIT)“ eine Person beschäftigen, welche die speziellen Anforderungen zwischen den Tier2-Rechenzentren in Deutschland, dem Tier1-Rechenzentrum in Karlsruhe und der ALICE Offline-Gruppe koordiniert. Diese Person hat ihren Arbeitsplatz am KIT und verhandelt alle besonderen GRID-Bedürfnisse des Experiments. Im Falle von ALICE wird diese Person zu 50% vom KIT finanziert. Die fehlenden 50% zur vollständigen Ausfinanzierung über die gesamte Förderperiode werden vom Antragsteller in Absprache mit den ALICE Projektpartnern im Rahmen dieser Projektperiode aus diesem Projekt finanziert und zu 100% an das KIT weitergeleitet.

### 2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Für die ALICE TPC Datenauslese wurde eine hardwarebasierte Datenvorverarbeitung in den Auslesekarten der FLP-Knoten geplant. Dazu müssen effiziente Algorithmen in Hardware auf den FPGA basierten CRU-Karten implementiert werden. Im ersten Schritt wurde versucht ein VHDL-Design von Dr. Sebastian Klevin in das CRU-Design zu integrieren. Wie sich herausstellte konnte das Design nicht in die bestehende Hardware integriert werden, da der Ressourcenverbrauch zu groß war. Daraufhin wurde das Design von Torsten Alt vollständig umgeschrieben und der Sorter und wenige Algorithmen, wie eine vereinfachte Zero-Suppression, mit der erforderlichen Ressourcen Einsparung in VHDL implementiert. Das war zeitintensiv und das Risiko des Scheiterns war sehr hoch. Aus dieser Erfahrung heraus wurde deshalb parallel dazu an einer Lösung gearbeitet, den Entwicklungsprozess zu vereinfachen und zu beschleunigen. Hierzu wurde mit der schrittweisen Implementierung und Integration bestehender Vorverarbeitungsalgorithmen mit Hilfe der High-Level-Synthese (HLS) begonnen, um die Machbarkeit dieser Vorgehensweise zu zeigen. Die Ergebnisse sind, dass sich die Hardwareressourcen mit Hilfe der Modern C++ Template Programmierung gut kontrollieren und ausnutzen kann. Aufbauend auf diesen Erkenntnissen wurde in dieser Förderperiode eine „Modern C++ HLS Datenfluss Template Library“ entwickelt, mit deren Hilfe sich komplexe Algorithmen als Deep Pipeline in Hardware implementieren lassen.



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

## 3 Planung und Ablauf des Vorhabens sowie Kooperation mit Dritten

Die Arbeiten wurden in Absprache mit den Arbeitsgruppen von Prof. Lindenstruth durchgeführt. Die nötigen Anforderungen der Template Library wurden mit der TPCU-CRU Gruppe am CERN abgestimmt. Der eingereichte Projektplan wurde eingehalten. Es wurde eine generische C++ Template Library für den Intel HLS Compiler entwickelt. Mit Hilfe der Library können Auslesealgorithmen als Datenflussgraph beschrieben und in Form einer Deep Pipeline auf Hardware implementiert werden. Diese wurde dann benutzt, um den nach Absprache der verschiedenen Arbeitsgruppen am CERN die erforderlichen Anforderungen für die Datenauslese zu verifizieren und typische Algorithmen zu implementieren.

## 4 Verwendung der Zuwendung (wichtigste Positionen des zahlenmäßigen Nachweises, z. B. Investitionen, Personalmittel)

Die Arbeite des WP6 wurden zu 100% über eine E13 Stelle finanziert. Die bewilligten Mittel für das GRID-Computing wurden an das KIT weitergeleitet.

## 5 Erzielte Ergebnisse mit Gegenüberstellung der vereinbarten Ziele

Basierend auf den Ergebnissen und Erfahrungen wurde zunächst eine HLS-Template Library entwickelt mit dem Fokus auf die Datenfluss-Programmierung. Dabei wurden die Anforderungen den Ressourcenverbrauch der Hardware im akzeptablen Limit zu halten erreicht. Auch der Entwicklungsaufwand hat sich dramatisch reduziert, da man den Algorithmus als C++ Code auf einer CPU emulieren kann, um so numerische Anforderungen zu verifizieren. Der gleiche Code wird dann ohne Änderung auf Hardware synthetisiert und reduziert dadurch mögliche Fehlerquellen enorm. Der Algorithmus wird als Datenflussgraph beschrieben und als Deep Pipeline massiv parallel ausgeführt. Das erlaubt die Datenvorverarbeitung kontinuierlicher Datenströme in Echtzeit. In der letzten Phase wurden verschiedene Algorithmen entworfen und getestet. Dabei wurden Fehler der Library identifiziert und behoben. Darüber hinaus wird aktuell daran gearbeitet die Library nach SYCL zu portieren.

## 6 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die geleistete Arbeit kann in zukünftigen Experimenten eingesetzt werden, um komplexe Algorithmen der Datenauslese von Detektoren nicht nur der Hochenergiephysik zu realisieren. Dabei wird der Algorithmus zunächst als C++ Programm auf CPU emuliert, um alle numerischen Anforderungen auszuarbeiten. Das reduziert Fehlerquellen und beschleunigt den Entwicklungsprozess.

## 7 Voraussichtlicher Nutzen, insbesondere Verwertbarkeit der Ergebnisse

Die Arbeiten des WP6 sind unerlässlich, um schnell und flexibel die immer komplexeren Algorithmen der Detektor-Auslese zu entwickeln. Es ist ein Beitrag die hohen Anforderungen der immer höher werdenden Datenraten gerecht zu werden. Die Ergebnisse können überall dort eingesetzt werden wo hohe Datenmengen in Echtzeit aus Detektoren ausgelesen und verarbeitet werden müssen, und ist nicht nur auf Detektoren der Hochenergiephysik beschränkt.

## 8 Während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordenen Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Keine Bekannt.



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

# 9 Erfolgte und geplante Veröffentlichungen der Ergebnisse

## 9.1 Referierte Publikationen (z. B. in Fachzeitschriften oder -büchern und referierte Konferenzproceedings)

Janson, Kebschull:
*HLS C++ Template Library for Detector Readout and Data-Preprocessing using FPGAs*
Virtual DPG Meeting, HK 11.5, SMuK 2021

Thomas Janson and Udo Kebschull:
*Detector Readout Algorithms and Data Flow Programming on FPGAs with Intel HLS*
DPG Virtual Spring Meeting, HK 37.2, Mainz 2022

Kebschull, Lindenstruth:
*Analysing the computational complexities of CERN's ALICE experiment*
https://www.innovationnewsnetwork.com/alice-epn-farm-cern-alice-experiment/20634/

Janson, Kebschull
*Data pre-processing with high-level-synthesis and dataflow programming using HLS C++ data-flow template library*
Nuclear Instruments and Methods in Physics Research Section A, Volume 1045, 167594, 2023, ISSN 0168-9002
https://doi.org/10.1016/j.nima.2022.167594

Janson, Kebschull
*Digital Signal Processing with FPGAs using Modern C++ and HLS*
DPG Spring Meeting, HK 26.4, Dresden 2023

Janson, Kebschull
*Modern C++17 data pre-processing HLS Dataflow Template Library*
IOP Publishing, Journal of Instrumentation, Volume 18, February 2023
https://doi.org/10.1088/1748-0221/18/02/C02050

*Janson, Kebschull*
*Modern C++ with SYCL as Multi Paradigm Programming Language for FPGA-Based Detector Readout*
*DPG Spring Meeting, HK 54.4, Gießen 2024*

## 9.2 Andere Veröffentlichungen (z. B. Konferenzbeiträge wie Vorträge und Poster, unreferierte Proceedings, Conference Notes)

## 9.3 Abschlussarbeiten (Bachelor, Master, Diplom, Staatsexamen, Promotion, Habilitation)



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

# Subprojekt Prof. Dr. Volker Lindenstruth:

## 1 Aufgabenstellung und Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Dieser Antrag ist im Kontext des ALICE at High Rate Projektes zu sehen. Für den LHC Run 3 benötigte ALICE wegen der deutlich höheren Datenraten und der weitgehend Trigger losen Auslese eine neue Computing-Infrastruktur. Die in diesem Projekt zusammengefassten Teilprojekte liefern einen zentralen Beitrag zum O2-EPN Projekt. Das ursprüngliche O2 Projekt, das im TDR beschrieben wurde, wurde während der Run-2 Periode in drei Teilprojekte zerlegt, das FLP Projekt, das die Daten von den 20.000 optischen Detektor Links empfängt, das PDP Projekt, das sich schwerpunktmäßig um die Ereignis Rekonstruktion und die Physik Analyse kümmert und das EPN Projekt, das sich um die HPC Farm von ALICE kümmert und diese betreibt. Hierzu gehören neben dem Schichtbetrieb auch die Entwicklung und Überwachung von Qualitätskontrollfunktionen. Gleichzeitig wird die EPN Farm als Grid Standort betrieben, was einige Weiterentwicklungen erforderte. Es gibt natürlich eine enge Zusammenarbeit zwischen den Projekten.

## 2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Im Vorfeld dieses Projektes wurden im Zusammenhang mit den vorangegangenen Förderungen erhebliche Entwicklungen für ALICE gemacht. So war der ALICE High Level Trigger (HLT) Cluster ein integraler Bestandteil des ALICE Run2 Betriebs. Seine Hauptaufgaben waren Datenkomprimierung, Ereignisauswahl, Analyse und Überwachung der Daten. Die Datenkomprimierung erlaubte eine mehr als Vervierfachung der Anzahl der Ereignisse, die zur Analyse gespeichert werden können. Die Online-Ereignisauswahl baute auf komplett rekonstruierten Ereignissen auf und erlaubt so komplexe Selektionsverfahren und eine sehr genaue Überwachung der Detektor-Performance. Diese Echtzeitanalyse und die Kalibrierung der Daten mittels Physikanalyse-Software ist durch modernste Softwareentwicklungen möglich gewesen und diente als Prototyp für Run3-Software. Der HLT wurde während des gesamten Run-2 sehr zuverlässig betrieben und hat eine sehr hohe Qualität der Ereignisrekonstruktion erreicht.

Nach Abschluss des Run-2 wurden erhebliche Vorbereitungsarbeiten für den Run-3 durchgeführt. Hierbei hat es auch eine Restrukturierung innerhalb von ALICE gegeben, dessen Ergebnis ist, dass wir das EPN Projekt leiten, das die on-line Rekonstruktionsfarm der EPN Server enthält. Die EPN Farm führt die gesamte synchrone (on-line, in Echtzeit) Datenanalyse durch und wird gleichzeitig für die weiterführende asynchrone (off-line) Analyse verwendet.

## 3 Planung und Ablauf des Vorhabens sowie Kooperation mit Dritten

Die Arbeiten wurden in enger Absprache mit den Arbeitsgruppen von Prof. Lindenstruth, Prof. Kebschull, den verschiedenen Arbeitsgruppen des ALICE $O^2$ Projektes, sowie in Kooperation mit den weiteren im Projektverbund beteiligten Arbeitsgruppen durchgeführt.

Es gab während der Projektlaufzeit ein Team bestehend aus vier Mitarbeitern mit denen mindestens wöchentliche Videokonferenzen abgehalten wurden. Typischerweise einmal monatlich ist der Projektleiter Prof. Lindenstruth zum CERN gefahren um an den Sitzungen des Technical Boards und des Management Boards teilzunehmen. In diesem Zusammenhang haben auch persönliche Treffen mit den am CERN stationierten Mitarbeitern stattgefunden. Gegen Ende jedes Jahres wurde ein Retreat in Frankfurt mit allen Projektteilnehmern abgehalten in dem die Erfolge und Probleme des letzten Jahres ausführlich besprochen wurden und die Prioritäten für den Winter Shutdown festgelegt wurden. In dieser Zeit wurden viele bekannte Probleme behoben und Verbesserungen vorgenommen. Insbesondere wurden hier auch die Wünsche der Runcoordination von ALICE umgesetzt. Besonders die Mitarbeiter am CERN haben gemeinsam mit der FLP Gruppe am CERN das Vorgehen und die Entscheidungen für die nächsten Schritte besprochen und geplant. Insbesondere während des Strahlbetriebes haben wöchentliche und teilweise sogar tägliche Meetings mit der ALICE Runcoordination stattgefunden.



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

## 4 Verwendung der Zuwendung (wichtigste Positionen des zahlenmäßigen Nachweises, z. B. Investitionen, Personalmittel)

Die genaue Verwendung der Zuwendung ist im zahlenmäßigen Verwendungsmachweis aufgelistet. Insgesamt wurden die Personalmittel zusammen mit den Eigenbeiträgen für die Entwicklungen und den Betrieb der EPN Farm im Projekt verwendet.

## 5 Erzielte Ergebnisse mit Gegenüberstellung der vereinbarten Ziele

Die Ergebnisse werden mit Bezug auf die einzelnen Unterprojekte im Folgenden dargestellt. Insgesamt wird festgestellt dass alle Meilensteine erreicht wurden.

**Aufbau und Betrieb der EPN Farm:**

Die heterogene Farm (nach der Erweiterung ab 2023) wurde während der gesamten Strahlzeit erfolgreich genutzt. Die Zahl der Bereitschaftsdienste ist auf weniger als 50 pro Jahr zurückgegangen, wobei es sich in den meisten Fällen entweder um Hardwareausfälle, Ausfälle des Orchestrierungssystems oder falsch erkannte Anrufe durch die Betreiber handelte. Es wurden keine EPN-bezogenen Vorfälle gemeldet, die sich auf die Online-Datenübernahme auswirkten, und nur ein Vorfall betraf synthetische Tests im Produktionssystem.

Aufgrund der Alterung hat die Farm einige ihrer Systeme wegen irreparabler Hardwareprobleme verloren. Dies waren aber weniger als 1 % der installierten Systeme. Die behebbaren Hardware-Probleme wurden alle während des Betriebs behoben, so dass eine ausreichende Rechenleistung für die Online-Datenübernahme gewährleistet war. Es gab über 50 Hardware-Austausche, wobei viele weitere Probleme entweder nicht reproduzierbar waren oder einen Austausch nicht rechtfertigten.

Das seit langem bestehende Problem mit dem AMD ROCm-Stack wurde in dieser Zeit nicht vollständig gelöst, und wir waren gezwungen, weiterhin eine maßgeschneiderte Anpassung des ROCm-Compilers und der Kernel-Module zu verwenden, was sich auf die Wartung und Leistung auswirkte.

Wir haben unsere Überwachungssysteme erheblich stabilisiert, indem wir alle Zeitreihen-Datenbankabfragen optimiert, unsere Wartungsverfahren verbessert, die gesamte Datenaufbewahrungszeit verkürzt und Backup-Instanzen aller Dienste eingerichtet haben. Darüber hinaus führten vereinfachte und übersichtlichere Dashboards für die Bediener dazu, dass deutlich weniger Anrufe von den Bedienern falsch identifiziert wurden.

Um den Betrieb weiter zu verbessern, wurden die Foreman-, DNS- und DHCC-Dienste komplett neu installiert, ein sterbender NFS-Server ersetzt und eine neue Version des Betriebssystems - Alma 9.x - vorbereitet und getestet. Das Betriebssystem-Upgrade ist leider aufgrund von Problemen mit dem AMD ROCm-Stack gegenwärtig noch blockiert.

Insgesamt ist die EPN Farm über die gesamte Projektzeit sehr zuverlässig gelaufen. Sie hat on-line Datenraten von bis zu 1,3 Terabyte pro Sekunde verarbeitet.

**Einsatz beim Strahlbetrieb:**

Im Jahr 2023 wurde die synchrone O2-Verarbeitung in der EPN-Farm vollständig eingesetzt und effektiv für die Sammlung von Pb-Pb-Daten mit hoher Intensität genutzt. Obwohl die meisten LHC-Fills mit etwa 25 kHz betrieben wurden, wurden einige Daten erfolgreich mit etwa 45 kHz gesammelt. Aufgrund der suboptimalen Effizienz der LHC-Injektoren konnten jedoch während der Pb-Pb-Kollisionen keine dauerhaften Daten mit hoher Wechselwirkungsrate aufgenommen werden. Während der Pb-Pb-Läufe behielt die EPN-Farm einen Spielraum von mehr als 10 % bei der Rechenkapazität im Vergleich zu den geschätzten Datenraten für den Pb-Pb-Betrieb mit 50 kHz.

Im Laufe des Jahres 2023 sammelte ALICE 38 PB an pp-Daten und 42 PB an Pb-Pb-Daten. Im Jahr 2024 wurden insgesamt 180 PB an pp-Daten aufgezeichnet (Pb-Pb-Daten aus dem Jahr 2024 sind nicht Gegenstand dieses Berichts). Was die integrierte Luminosität betrifft, so erreichte ALICE 1,53



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

$nb^{-1}$ bei Pb-Pb-Kollisionen und 14,5 $pb^{-1}$ bei pp-Kollisionen, mit einer Gesamteffizienz von 76 %. Bemerkenswert ist, dass der Betrieb der EPN-Farm nicht zu den Ineffizienzen am Ende des Laufs beitrug; die meisten Verluste in der Laufeffizienz wurden auf Probleme mit Detektoren oder anderen zentralen Systemen zurückgeführt.

Trotz dieser Herausforderungen war die Menge der gesammelten Daten beträchtlich und entsprach den kombinierten Statistiken von Lauf 1 und Lauf 2. Um diesen Zustrom zu bewältigen, wurde eine Strategie eingeführt, um nur die relevantesten Ereignisse auf der Grundlage der „bunch crossings" des LHC Strahls zu speichern. Dazu gehörte das Überfliegen komprimierter Zeitrahmen, um nur 3-4,5 % der ursprünglichen Daten auf der Festplatte zu speichern. Dieser Ansatz war entscheidend, um zu verhindern, dass die 100 PB Festplattenpufferkapazitäten von CERN IT überschritten werden, und gewährleistete eine ununterbrochene Datenerfassung, wodurch die Nachhaltigkeit des ALICE-Betriebs sichergestellt wurde.

**übergreifende Qualitätssicherung:**

Ein großer Fortschritt bei der Online-Analyse wäre die Rekonstruktion von kurzlebigen Teilchen in der asynchronen Rekonstruktionsphase von ALICE mit Hilfe der Kalman-Filter-Software (KF) zur Teilchenidentifizierung. Die Software ist perfekt für die Implementierung auf den EPN-GPUs geeignet, da sie parallelisierbar ist. Neben der bestehenden Infrastruktur für die Qualitätskontrolle würde ein solches Tool weitere Gegenkontrollen in Bezug auf die Einzelspurinformationen ermöglichen. Das endgültige Ziel wäre jedoch die Extraktion von Observablen aus dem rekonstruierten Teilchen und deren Vergleich mit bestehenden theoretischen Vorhersagen, um eine robuste Überprüfung der Datenqualität in der asynchronen Phase zu ermöglichen.

Ein generisches Testpaket wurde erweitert, um die Leistung der KFParticle-Software anhand von MC-Simulationen zu überprüfen. Darüber hinaus wurde eine Aufgabe, die sich auf schwere Teilchen konzentriert, als Testrahmen in der O2-Software entwickelt. Sie bietet einen Weg zur Analyse der topologischen Selektionen auf dem Zerfallsteilchen. Die Tests wurden mit Run-3-Simulationen und -Daten durchgeführt, und es wurde eine ähnliche Leistung wie mit der Standardsoftware erzielt. Bei detaillierteren Untersuchungen mit der Standardsoftware für die kurzlebige Rekonstruktion von ALICE (zusammen mit KFParticle) wurden jedoch einige Trends in den anfänglichen asynchron rekonstruierten Daten beobachtet, wie z. B. eine Verbreiterung der Spitzenwerte der invarianten Masse, Unstimmigkeiten zwischen den Erträgen der Teilchen und ihren konjugierten Ladungen und Asymmetrien der DCAs zwischen den Zerfallsprodukten. Diese sind das Ergebnis einer Verschlechterung der Spurimpulsauflösung im Vergleich zu Run 2, die höchstwahrscheinlich auf größere Raumladungsverzerrungen in der TPC infolge der höheren Raten zurückzuführen ist. Die Entwicklungen bei der TPC-Fehlerparametrisierung und den TPC-Verzerrungskorrekturen führen zu erheblichen Verbesserungen. Zuvor waren für die Kalibrierung mehrere Iterationen über die Daten erforderlich. Jetzt wurden einige Aufgaben zusammengefasst, und z. B. können Driftgeschwindigkeits- und Korrekturkarten in einem Durchgang extrahiert werden. Die Arbeiten werden fortgesetzt, wobei sich die kurzfristigen Pläne auf Korrekturen und Kalibrierungen in einer einzigen Iteration konzentrieren, um schließlich das ursprüngliche Ziel der Kalibrierung in der synchronen Phase zu erreichen.

## 6 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die Arbeiten aller Arbeitspakete waren für den Betrieb in Run 3 notwendig. Ohne die EPN Farm ist der Betrieb von ALICE nicht möglich. Auch die Arbeiten zur Qualitätssicherung sind insbesondere wichtig, um einen Reibungslosen Betrieb der Detektoren zu gewährleisten.

## 7 Voraussichtlicher Nutzen, insbesondere Verwertbarkeit der Ergebnisse

Die Arbeiten aller Arbeitspakete sind für den Betrieb von ALICE in Run 3 zwingend notwendig. Ohne die funktionierende EPN Farm hätte ALICE nicht annähernd die Datenraten aufzeichnen können. Die ALICE EPN Farm ist die einzige on-line Farm bei der nahezu 100% der Algorithmen on-line auf GPUs



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

laufen, wodurch erhebliche Kosten eingespart wurden. Andere Experimente am CERN bereiten nun ebenfalls diesen Wechsel vor. Die ALICE EPN Farm ist auch ein Rollenmodell für die entsprechende on-line Farm, den First Level Event Selector (FLES) des CBM Experiments bei FAIR.

## 8 Während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordenen Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Es gibt eine Zusammenarbeit zwischen ALICE und FAIR. Die GSI ist auch Mitglied der ALICE Kollaboration. Nachdem ALICE im Run 3 vor FAIR den triggerlosen Betrieb genommen hat, wurden wertvolle Erfahrungen für CBM gesammelt.

## 9 Erfolgte und geplante Veröffentlichungen der Ergebnisse

Es sind im Berichtzeitraum etwa 270 Publikationen der ALICE Kollaboration mit den Projektleitern als Koautoren im Förderzeitraum entstanden. Hier werden nur einige Auszüge daraus gelistet, die für dieses Teilprojekt besonders spezifisch sind.

### 9.1 Referierte Publikationen (z. B. in Fachzeitschriften oder -büchern und referierte Konferenzproceedings)

ALICE UPS ITS GAME FOR SUSTAINABLE COMPUTING, Volker Lindenstruth for the ALICE Collaboration, CERN Courier September/Oktober 2023

### 9.2 Andere Veröffentlichungen (z. B. Konferenzbeiträge wie Vorträge und Poster, unreferierte Proceedings, Conference Notes)

### 9.3 Abschlussarbeiten (Bachelor, Master, Diplom, Staatsexamen, Promotion, Habilitation)

**Bachelor**

Kraynykov, Vadim: Accelerating FEBID Simulaton on Heterogenous Computing Sys-tems using Py-OpenCL

**Master**

Anna Bartsch (Physik): Initialization of Best-fit Splines for the Calibration of the ALICE TPC

Abu-Ayyad, Marwa: Vorhersage des Migrationstyps von Lymphomzellen mithilfe von convolutional neural Networks (CNNs)

Cezik, Michael Alper: Quark-Gluon Jet Discrimination using Jet Images and deep neural Networks

**Dissertation**

Johannes Lehrbach: High Throughput Computing Infrastructure for the ALICE EPN Online Processing



Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

# Kurzbericht

- öffentlich -

| | |
|---|---|
| Zuwendungsempfänger: | Goethe-Universität Frankfurt am Main |
| Projektleitung: | Prof. Dr. Volker Lindenstruth |
| Verbund: | 05P21RFCA2: Verbundprojekt 05P2021 (ErUM-FSP T01)- Run 3 von ALICE am LHC: Fertigstellung und Inbetriebnahme der EPN-Farm” |
| Thema: | Arbeitspakete – Prof. Dr. U. Kebschull<br>o Entwicklung schneller Datenvorverarbeitung mittels High-Level-Synthese im FLP mit Schwerpunkt auf die Time Projection Chamber (WP6, TPC Readout)<br><br>Arbeitspakete – Prof. Dr. V. Lindenstruth<br>o Aufbau und Betrieb der EPN Farm<br>o Einsatz beim Strahlbetrieb<br>o übergreifende Qualitätssicherung |

## 1. Ziel und Inhalt des Projektes

Die Arbeiten der Antragssteller wurden im Kontext der im ErUM-FSP T01 organisierten deutschen ALICE Universitätsgruppen in Frankfurt, Münster, Heidelberg, München, Bonn, Bielefeld und Tü-bingen sowie der GSI durchgeführt. Sie stellen einen erheblichen Beitrag zum ALICE Experiment dar. Hier sind insbesondere die TPC, der TRD und die EPN-Rechenfarm zu nennen. Die Projektlei-tung dieser drei Teile von ALICE liegt bei den deutschen Gruppen. Darüber hinaus gab es eine intensive Zusammenarbeit mit der FLP und der PDP Gruppe am CERN im Kontext des Rechner-betriebes und besonders im Bereich der Entwicklung der schnellen und effizienten Ereignisrekon-struktionssoftware auf GPUs. Die EPN-Rechenfarm ist bezüglich des sehr konsequenten Einsat-zes von GPU-Rechenbeschleunigern am CERN führend. Im Kontext der Validierung der verschie-denen Softwarekomponenten der ALICE Subdetektoren wurde mit den entsprechenden Verant-wortlichen der Detektorgruppen, insbesondere der TPC, zusammengearbeitet und Hilfestellung bei der Softwareentwicklung geleistet. Es besteht eine Validierungsinfrastruktur die alle neue Software zu durchlaufen hat um sicherzustellen, dass die Software möglichst fehlerfrei ist.

Das ALICE Experiment am Large Hadron Collider des CERN wurde während des Run 3 betrieben und weiter entwickIt. Die ALICE Event Processing Farm (EPN) ist eine online Farm, die die Daten des Experiments in Echtzeit analysiert und komprimiert. Hierbei werden die Daten des Detektors mit einer Rate von knapp einem Terabyte pro Sekunde verarbeitet. Eine maximale Verarbeitungsrate von 1,3 Terabyte pro Sekunde konnten demonstriert werden. Hierbei wird eine volle Ereignisrekonstruktion durchgeführt. Ein wesentliches Element dieser on-line high-throughput HPC Farm ist der konsequente Einsatz von Graphikkarten als Hardware Beschleuniger um die Kosten zu reduzieren. Mittlerweile sind fast 100% der Software des synchronen Rechnens (online) auf Graphikkarten lauffähig. Insgesamt kann derselbe Quellcode sowohl auf verschiedenen GPUs aber auch auf CPUs

laufen. Die EPN Farm implementiert 2800 AMD MI50/MI100 Server GPUs, 24.640 physikalische AMD CPU Kerne, 200 TB Hauptspeicher und ein 100 Gb/s InfiniBand Netzwerk. Die Farm besteht aus 350 individuellen Servern. Die EPN Farm wird auch für das asynchrone (offline) Rechnen verwendet. Wesentliche Entwicklungen waren der TPC Readout, die Monitoring Umgebung und die Qualitätskontrollinfrastruktur für die Datenaufnahme. Ein zentraler Baustein der EPN Farm ist die Daten- und Lastverteilung, die die gesamten Datenströme in der EPN Farm steuert. Sie wurde während der Förderperiode zusammen mit aller notwendigen Software fertiggestellt und ist seitdem im Einsatz.

## 2. Ablauf und Ergebnisse des Vorhabens

Basierend auf den Ergebnissen und Erfahrungen wurde zunächst eine HLS-Template Library entwickelt mit dem Fokus auf die Datenfluss-Programmierung. Dabei wurden die Anforderungen den Ressourcenverbrauch der Hardware im akzeptablen Limit zu halten erreicht. Auch der Entwicklungsaufwand hat sich dramatisch reduziert, da man den Algorithmus als C++ Code auf einer CPU emulieren kann, um so numerische Anforderungen zu verifizieren. Der gleiche Code wird dann ohne Änderung auf Hardware synthetisiert und reduziert dadurch mögliche Fehlerquellen enorm. Der Algorithmus wird als Datenflussgraph beschrieben und als Deep Pipeline massiv parallel ausgeführt. Das erlaubt die Datenvorverarbeitung kontinuierlicher Datenströme in Echtzeit. In der letzten Phase wurden verschiedene Algorithmen entworfen und getestet. Dabei wurden Fehler der Library identifiziert und behoben. Darüber hinaus wird aktuell daran gearbeitet die Library nach SYCL zu portieren.

Die heterogene EPN-Farm (nach der Erweiterung ab 2023) wurde während der gesamten Strahlzeit erfolgreich genutzt. Die Zahl der Bereitschaftsdienste ist auf weniger als 50 pro Jahr zurückgegangen, wobei es sich in den meisten Fällen entweder um Hardwareausfälle, Ausfälle des Orchestrierungssystems oder falsch erkannte Anrufe durch die Betreiber handelte. Es wurden keine EPN-bezogenen Vorfälle gemeldet, die sich auf die Online-Datenübernahme auswirkten, und nur ein Vorfall betraf synthetische Tests im Produktionssystem. Insgesamt ist die EPN Farm über die gesamte Projektzeit sehr zuverlässig gelaufen. Sie hat on-line Datenraten von bis zu 1,3 Terabyte pro Sekunde verarbeitet.

Im Laufe des Jahres 2023 sammelte ALICE 38 PB an pp-Daten und 42 PB an Pb-Pb-Daten. Im Jahr 2024 wurden insgesamt 180 PB an pp-Daten aufgezeichnet (Pb-Pb-Daten aus dem Jahr 2024 sind nicht Gegenstand dieses Berichts). Was die integrierte Luminosität betrifft, so erreichte ALICE 1,53 $nb^{-1}$ bei Pb-Pb-Kollisionen und 14,5 $pb^{-1}$ bei pp-Kollisionen, mit einer Gesamteffizienz von 76 %. Bemerkenswert ist, dass der Betrieb der EPN-Farm nicht zu den Ineffizienzen am Ende des Laufs beitrug; die meisten Verluste in der Laufeffizienz wurden auf Probleme mit Detektoren oder anderen zentralen Systemen zurückgeführt.

Ein generisches Testpaket wurde erweitert, um die Leistung der KFParticle-Software anhand von MC-Simulationen zu überprüfen. Darüber hinaus wurde eine Aufgabe, die sich auf schwere Teilchen konzentriert, als Testrahmen in der O2-Software entwickelt. Sie bietet einen Weg zur Analyse der topologischen Selektionen auf dem Zerfallsteilchen. Die Entwicklungen bei der TPC-Fehlerparametrisierung und den TPC-Verzerrungskorrekturen führen zu erheblichen Verbesserungen, z.B. können Driftgeschwindigkeits- und Korrekturkarten in einem Durchgang extrahiert werden. Die Arbeiten werden fortgesetzt, wobei sich die kurzfristigen Pläne auf Korrekturen und Kalibrierungen in einer einzigen Iteration konzentrieren, um schließlich das ursprüngliche Ziel der Kalibrierung in der synchronen Phase zu erreichen.

## 3. Darstellung der wesentlichen Ergebnisse und deren konkreter Nutzen sowie ggf. die Zusammenarbeit mit anderen Forschungseinrichtungen

Die hier durchgeführten Arbeiten ermöglichen erst den Betrieb von ALICE da in Run 3 die online Rekonstruktion der Daten erforderlich ist. Die Arbeiten wurden abgeschlossen und das Experiment befand sich zu Projektende im Strahlbetrieb.

Ein Großteil der Arbeiten dieses Projektes sind unter anderem für FAIR am Helmholtzhentrum GSI übertragbar.

Durch den Einsatz moderner Technologien wie InfiniBand und GPGPUs konnten erhebliche finanzielle Mittel eingespart werden, da die Kosten einer nur auf CPUs basierenden EPN Farm die bestehenden Kosten um einen Faktor sieben erhöht hätten.