Kurzbericht

- öffentlich -

Zuwendungsempfänger:	Zuse-Institut Berlin (ZIB)
Projektleitung:	Dr. Florian Schintke
Verbund:	Verbundprojekt 05P2021 (ErUM-FSP T06) - Aufbau von CBM bei FAIR
Thema:	CBM/FLES: Elastizität und heterogene Datenvolumina in FLESnet
Förderzeitraum:	1.7.2021 – 31.12.2024

1. Ziel und Inhalt des Projektes

In den Experimenten am FAIR Teilchenbeschleuniger wird eine fortschrittliche Ausleseelektronik eingesetzt, die direkt in das Detektorsystem integriert ist. Eine Vielzahl von Detektoren sammelt Messdaten und versieht sie mit Zeitstempeln. Diese Datenströme werden anschließend zusammengefasst und für weitere Verarbeitungsprozesse an verschiedenen Schnittstellen zum FLES-System (first-level event selector) weitergeleitet. Im Gegensatz zu älteren Ansätzen erfolgt hier eine direkte Selektion von Ereignissen in Echtzeit. Hierzu werden die Messdaten in spezifische Zeitfenster oder Zeitscheiben organisiert, analysiert und bei Bedarf an nachgelagerte Analysestufen weitergeleitet.

In diesem Projekt sollte das Flesnet System, das konsistente Zeitscheiben mit ihren dazu gehörenden Daten aus den Messdatenströmen auf Analyserechnern zusammenfügt und bereitstellt, erweitert werden, um eine dynamische Veränderung der Rechneranzahl während der Laufzeit zu ermöglichen (Elastizität) und den Umgang mit heterogenen Datenvolumina der einzelnen Messkanäle zu verbessern. Außerdem sollte die Skalierbarkeit von Flesnet untersucht und verbessert werden sowie die entwickelten Arbeiten im Rahmen des Laboraufbaus mini-CBM integriert werden.

2. Ablauf und Ergebnisse des Vorhabens

Das Projekt konzentrierte sich auf die technische Entwicklung und Verbesserung von Flesnet, das eine zentrale Rolle im CBM-Experiment an GSI/FAIR einnimmt. Flesnet aggregiert Messdatenströme und leitet sie in Zeitabschnitten (Zeitscheiben) an Analysesysteme weiter – eine Funktion, die für die spätere physikalische Auswertung von Kollisionsevents unerlässlich ist.

Im Mittelpunkt der Entwicklung stand die Stärkung der Fähigkeit von Flesnet, Fehler zu verarbeiten (Elastizität) und Daten aus verschiedenen Quellen mit heterogenen Datenmengen zu verarbeiten. Hierzu wurde Flesnet in einem InfiniBand-Netzwerkcluster mit einer modernen Linux-Umgebung und RDMA-basierter Kommunikation via libfabric am ZIB eingesetzt und erweitert. Parallel dazu umfasste die kontinuierliche Weiterentwicklung von Flesnet

Anpassungen an sich verändernde Bibliotheks-Versionen (insbesondere bei libfabric), um Stabilität und Kompatibilität weiter sicherzustellen [3].

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

Das CBM-Experiment am GSI/Fair benötigt Flesnet als wichtige Komponente auf dem Datenverarbeitungspfad. Flesnet aggregiert und leitet die Messdatenströme in Zeitscheiben zu Analyserechnern. Die im Projekt adressierten Ziele der Unterstützung von Elastizität und heterogenen Datenvolumina für Flesnet sind dabei im geplanten praktischen Einsatz des Systems wichtige Eigenschaften. Durch die entwickelten Anpassungen bezüglich dieser Aspekte im Rahmen dieses Projekts konnte die Unterstützung des Systems für Elastizität und heterogene Datenvolumina deutlich verbessert werde, was auch unsere praktische Evaluation zeigt.

Im Bereich Skalierbarkeit sind unsere praktischen Messergebnisse auf einem großen System mit insgesamt 96 Rechnern (48 Eingangs- und 48 Prozessierungsrechner) [2], die eine stabile aggregierte Datenrate von bis zu 340 GB/s erreicht haben, auch Teil des ersten technische Entwurfsbericht „Technical Design Report for the CBM – Online Systems – Part I“ des CBM-Projekts für zentrale Systeme geworden. Dieser Bericht wurde vom Expertenausschuss Experimente (ECE) fachlich begutachtet und auf dessen Empfehlung von FAIR im Juli 2023 genehmigt und veröffentlicht [1] (DOI: 10.15120/GSI-2023-00739).

Die entwickelten Funktionalitäten wurden auch erfolgreich auf der mini-CBM-Infrastruktur mit Integrationstests überprüft. Die Ergebnisse des Projekts tragen zur Gesamtbereitschaft des CBM-Experiments bei und gewährleisten eine effiziente und zuverlässige Datenverarbeitung. Die Software selbst ist Open Source, was eine breitere Akzeptanz und weitere Entwicklung innerhalb der CBM-Kollaboration und darüber hinaus fördert.

Der tatsächliche Aufbau des CBM-Experiments und des Teilchenbeschleuniger SIS100 schreiten stetig voran, wie auch ein Blick auf die Baustelle am GSI/FAIR verrät. Für den tatsächlichen Einsatz im Experimentbetrieb sind weitere Untersuchungen, Stabilisierungen und Anpassungen an neue Gegebenheiten von Flesnet notwendig, die in einer weiteren Projektförderphase adressiert werden sollten.

Das Projekt wurde in Abstimmung mit der CBM-Collaboration, insbesondere dem CBM/FLES Projekt und der Forschungsgruppe von Prof. Lindenstruth (Goethe-Universität Frankfurt a. M.) durchgeführt.

Literaturreferenzen

[1] CBM Collaboration. *Technical Design Report for the CBM Online Systems – Part I, DAQ and FLES Entry Stage.* J. d. Cuveland, D. Emschermann, V. Friese, I. Fröhlich, P. Gasik, D. Hutter, W. Müller, and C. Sturm (editors). Technical Report, Darmstadt, 2023. FAIR Technical Design Report. URL: https://repository.gsi.de/record/340597, doi:10.15120/GSI-2023-00739.

[2] F. Salem, F. Schintke. *Large-Scale Performance of the Data-Flow Scheduler (DFS) and FLESnet.* CBM Progress Report 2021, pp. 170-171, 2022, doi: 10.15120/GSI-2022-00599.

[3] F. Schintke, N. Greve, J. de Cuveland. *Recent Flesnet developments and work in progress.* CBM Progress Report 2023, p. 132, 2024, doi: 10.15120/GSI-2024-00765.

Schlussbericht

Zuwendungsempfänger:	Zuse-Institut Berlin (ZIB)
Projektleitung:	Dr. Florian Schintke
Verbund:	Verbundprojekt 05P2021 (ErUM-FSP T06) - Aufbau von CBM bei FAIR
Thema:	CBM/FLES: Elastizität und heterogene Datenvolumina in FLESnet
Förderzeitraum:	1.7.2021 – 31.12.2024

Zusammenfassung

Das Projekt konzentrierte sich auf die technische Entwicklung und Verbesserung von Flesnet, das eine zentrale Rolle im CBM-Experiment an GSI/FAIR einnimmt. Flesnet aggregiert Messdatenströme und leitet sie in Zeitabschnitten (Zeitscheiben) an Analysesysteme weiter – eine Funktion, die für die spätere physikalische Auswertung von Kollisionsevents unerlässlich ist.

Im Mittelpunkt der Entwicklung stand die Stärkung der Fähigkeit von Flesnet, Fehler zu verarbeiten (Elastizität) und Daten aus verschiedenen Quellen mit heterogenen Datenmengen zu verarbeiten. Hierzu wurde Flesnet in einem InfiniBand-Netzwerkcluster mit einer modernen Linux-Umgebung und RDMA-basierter Kommunikation via libfabric am ZIB eingesetzt und erweitert. Parallel dazu umfasste die kontinuierliche Weiterentwicklung von Flesnet Anpassungen an sich verändernde Bibliotheks-Versionen (insbesondere bei libfabric), um Stabilität und Kompatibilität weiter sicherzustellen.

Im Bereich Skalierbarkeit sind unsere praktischen Messergebnisse auf einem großen System mit insgesamt 96 Rechnern (48 Eingangs- und 48 Prozessierungsrechner), die eine stabile aggregierte Datenrate von bis zu 340 GB/s erreicht haben, auch Teil des ersten technische Entwurfsbericht „Technical Design Report for the CBM – Online Systems – Part I“ des CBM-Projekts für zentrale Systeme geworden. Dieser Bericht wurde vom Expertenausschuss Experimente (ECE) fachlich begutachtet und auf dessen Empfehlung von FAIR im Juli 2023 genehmigt und veröffentlicht [2] (DOI: 10.15120/GSI-2023-00739).

Die entwickelten Funktionalitäten wurden auch erfolgreich auf der mini-CBM-Infrastruktur mit Integrationstests überprüft. Die Ergebnisse des Projekts tragen zur Gesamtbereitschaft des CBM-Experiments bei und gewährleisten eine effiziente und zuverlässige Datenverarbeitung. Die Software selbst ist Open Source, was eine breitere Akzeptanz und weitere Entwicklung innerhalb der CBM-Kollaboration und darüber hinaus fördert.

Für zukünftige Entwicklungsphasen ist davon auszugehen, dass weitere Feinabstimmungen erforderlich sein werden, um Flesnet optimal an die sich weiterentwickelnden Anforderungen des CBM-Experiments anzupassen. Die intensive Zusammenarbeit innerhalb der CBM-Gemeinschaft und die Veröffentlichung der Ergebnisse in technischen Fachpublikationen tragen dazu bei, den Forschungsstand fortlaufend zu aktualisieren und neue Ansätze im Bereich skalierbarer, latenzarmer Datenverarbeitungssysteme für Hochenergiephysikexperimente weiterzuentwickeln.

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

1 Aufgabenstellung und Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Derzeit findet der Aufbau des neuen Teilchenbeschleunigers (SIS100) am GSI/FAIR in Darmstadt statt. Mit einem großen, innovativen Detektorsystem, wird das „Compressed Baryonic Matter“-Experiment (CBM) diesen Teilchenbeschleuniger für seine Messungen und wissenschaftlichen Untersuchungen nutzen. Auch dieses Experiment befindet sich momentan im Aufbau. Eine kleine „Laborversion“ (miniCBM oder kurz mCBM) mit einzelnen Stellvertretern der später verwendeten Detektorelemente der Subsysteme ist bereits regelmäßig an der GSI im Einsatz, wird kontinuierlich erweitert und angepasst, um sowohl die einzelnen Komponenten aber auch ihre Gesamtintegration zu testen und wichtige Praxiserfahrungen auf allen Ebenen zu sammeln. Die Messdaten werden bei CBM im Detektor erfasst, digitalisiert und über mehrere hundert Glasfaserverbindungen als Datenströme aus dem Detektor herausgeleitet. Die Daten werden vom sogenannten „First-Level Event Selector“-System (FLES) entgegengenommen und für die Datenanalyse vorbereitet und zusammengestellt. Die Datenströme werden von PCIe-FPGA Boards in zahlreichen Computern empfangen und in deren Hauptspeicher geschrieben. Von dort werden die Datenströme von einem Softwaresystem „Flesnet“ gelesen und so umorganisiert, dass jeweils Zeitscheiben (Timeslices) des Datenstroms mit allen Messdaten dieses Zeitraumes entstehen. Insgesamt wird eine Datenrate von einigen Terabyte pro Sekunde erwartet. Unterschiedliche Zeitscheiben werden dabei an verschiedene Rechner verteilt, um eine Überlastung eines einzelnen Rechners zu vermeiden.

Die Aufgabe dieses Projektes war es, den bisherigen Flesnet-Prototypen zu erweitern und anzupassen, um Flesnet weiter auf den produktiven Einsatz vorzubereiten. Dazu sollten wichtige Fähigkeiten hinzugefügt werden. Je nach Rechenbedarf und Datenaufkommen soll die Größe des Flesnet-Systems im Betrieb variiert werden (Elastizität), ohne dass jeweils das ganze Flesnet-System umkonfiguriert und neu gestartet werden muss. Hierbei müssen dann die Datenströme und die Datenverteilung — synchronisiert und abgestimmt über alle Knoten — im laufenden Betrieb neu organisiert werden.

Außerdem sollte Flesnet mit heterogenen Timeslice-Beiträgen untersucht und die erreichte aggregierte Bandbreite evaluiert werden. Hierzu war der Lastausgleich, das Puffermanagement und die koordinierte Netzwerknutzung weiter zu verfeinern und zu optimieren. Für die Evaluation sollten neben künstlichen Größenverteilungen der Daten auch abgespielte Daten des mini-CBM Systems kommen.

Gleichzeitig sollte Flesnet auf größeren Systemen evaluiert und untersucht werden, um näher an die Zieldatenrate von über 1 TByte/s heran zu kommen, das Skalierungsverhalten zu beobachten und entsprechende Verbesserungen vorzunehmen. Die Entwicklungen sollten auch in das mini-CBM System integriert und dort eingesetzt werden. Unsere gesammelten Erfahrungen und neuen Erkenntnisse sollten aktiv in den geplanten „Technical Design Report“, der unter anderem das FLES-System beschreibt und zur technisch-/wissenschaftlichen Begutachtung eingereicht wurde, eingebracht werden.

2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Das CBM-Experiment arbeitet, ähnlich wie das ALICE Experiment am CERN [1], mit einer selbst getriggerten Ausleseelektronik und leitet dafür alle gemessenen Daten mit Zeitstempeln an den First-Level Event Selector (FLES), der durch Online-Datenanalyse und Eventselektion die eigentliche Triggerfunktion implementiert. Hierfür wird für jeweils kurze Zeitabschnitte ein Gesamtbild der Beobachtungsdaten benötigt. Das Flesnet System hat daher die Aufgabe die über viele Eingangskanäle verteilten Messdatenströme in Zeitscheiben (Timeslices) auf Prozessierungsknoten zusammen zu führen und zur Verfügung zu stellen. Die Datenflüsse in Flesnet bilden aufgrund ihrer hohen Bandbreite, zeitlichen Abhängigkeiten und großen Anzahl von gleichzeitig zu bedienenden Eingangs- und Ausgangskanälen von einigen hundert eine große Herausforderung für die Koordination durch Flesnet und das darunter liegende tatsächliche Kommunikationsnetzwerk. Es besteht sowohl die Gefahr der Überlastung einzelner

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.

Datenempfänger (endpoint congestion), als auch der gesamten Netzwerkinfrastruktur (inner congestion), was jeweils zu drastischen Leistungseinbußen des Gesamtsystems führt. Solche Lastsituationen werden aktuell auch im Bereich des Supercomputings vermehrt als Herausforderung erkannt. Verschiedene Netzwerktechnologien, die zum Teil einen direkten Nachrichtenaustauch aber auch den oft effizienteren *remote direct memory access (RDMA)* nutzen, sind unterschiedlich gut darauf vorbereitet [4] und versuchen mit Techniken wie adaptivem Routing einem breiten Spektrum von Anwendungen gute Leistung zu bieten. Wie gut den einzelnen Netzwerken dies in den verschiedenen Congestion-Szenarien gelingt war bisher kaum abschätzbar und quantifizierbar. Mit dem GPCNeT Benchmark [3], wurde hierfür kürzlich eine interessante Vergleichsmethodik vorgeschlagen.

Auf dem Hochleistungsrechner HLRN-III am ZIB wurde Flesnet erfolgreich mit 384 Knoten ausgeführt. Durch den am ZIB für Flesnet entwickelten *Data-Flow Scheduler (DFS)* verbesserte sich die erreichbare aggregierte Bandbreite um bis zu 50 %, während die Zeit zwischen Ankunft des ersten und letzten Datenbeitrags der Input-Knoten für eine Timeslice um mehr als den Faktor 30 bei wesentlich geringerer Varianz reduziert werden konnte [5]. Möglich ist dies durch eine wesentlich reduzierte und ausgeglichenere Nutzung der Empfangspuffer in den Prozessierungsknoten, so dass auch in größeren Systemen die Puffer kleiner dimensioniert werden können [6].

3 Planung und Ablauf des Vorhabens sowie Kooperation mit Dritten

Die Arbeiten wurden in drei Arbeitspakete strukturiert, deren Planung und Ablauf in den folgenden Unterabschnitten dargestellt wird.

AP 1: Elastizität in Flesnet

Das bisherige Flesnet-System ist weitgehend auf eine statische Anzahl an Eingangs- und Analyse-Rechnern ausgelegt. Beispielsweise werden erst Timeslices gebaut, wenn alle Eingangsknoten verfügbar sind, da Timeslices mit partiellen Daten für die weitere physikalische Analyse unbrauchbar sind. Beim Start des Systems muss Flesnet daher synchronisiert auf den Eingangs- und Analyse-Rechnern gestartet werden. Für eine Veränderung der teilnehmenden Rechner muss ein kompletter Neustart des Flesnet-Systems erfolgen. Um diesen Nachteil zu umgehen, sollte Flesnet sowohl flexibler mit dynamisch verfügbaren Eingangs- als auch Analyse-Rechnern umgehen können (Elastizität). Im praktischen Betrieb des Flesnet-Systems koordinieren sich dann die bereits gestarteten Systeme untereinander und entscheiden gemeinsam ab welchem Zeitpunkt Timeslices gebaut werden sollen. Dies erlaubt eine frühe Inspektion des Detektorzustands bereits beim Hochfahren und Justieren, sobald partielle Timeslices erzeugt werden, so dass noch fehlende Kanäle und die richtige Ausrichtung des Teilchenstrahls frühzeitig beurteilt werden können.

Um das Bauen von Timeslices durch Flesnet einerseits überprüfen zu können und andererseits das Einbauen von Fehlern durch unsere Weiterentwicklungen zu verhindern beziehungsweise frühzeitig zu erkennen, haben wir einen sogenannten „TimesliceValidator“ implementiert, der eine Menge gegebener Microslice-Archives mit einer Menge gegebener Timeslice-Archives vergleicht und inhaltlich überprüft, ob alle Daten der Microslices sich auch in den Timeslice-Daten wiederfinden und umgekehrt. Dadurch lässt sich beispielsweise erkennen ob Timeslices fehlen, korrupt sind oder verworfen wurden. Durch diese Konsistenzprüfung können wir beispielsweise auch erkennen, wenn Eingangsknoten bereits Daten verteilt haben, die aufgrund der Startreihenfolge noch von keinem Empfangsrechner zu Timeslices verarbeitet werden konnten.

Eine neu entwickelte Softwarearchitektur nutzt das Konzept eines zentralen Managers (Central Manager), der die Koordinierung zwischen Sendern und Empfängern unterstützen soll und Entscheidungsprozesse so einfacher konsistent durchführen kann. Weiterhin wird auf effiziente Kommunikation in modernen Netzwerken geachtet (remote direct memory access) und als Abstraktionsschicht für solche Netzwerke Libfabric für die Implementation eingesetzt. Unsere Entwicklungstests zeigen, dass

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.

erfolgreich bei laufendem Datentransfer sowohl Knoten hinzugefügt, als auch entfernt werden können und das System sich dynamisch den neuen Gegebenheiten anpasst und die Last entsprechend anders verteilt.

Für die Elastizität des Flesnet-Systems wurden verschiedene Verfahren entwickelt und untersucht. Abbildung 1 zeigt die entsprechende Software-Architektur. Eine globale Sicht über die Menge der Eingangsknoten und deren Zustand erlaubt koordinierte Entscheidungen, um ein robustes und konsistentes Systemverhalten zu erreichen. Beim dynamischen Hinzufügen von Analyse-Rechnern werden die Datenströme neu organisiert und die Lastverteilung entsprechend angepasst. Der Einfluss einer geteilten Ressourcennutzung durch Eingangs- und Analyse-Knoten auf einem System wurde experimentell für Flesnet untersucht und bei der Last- und Datenverteilung berücksichtigt.

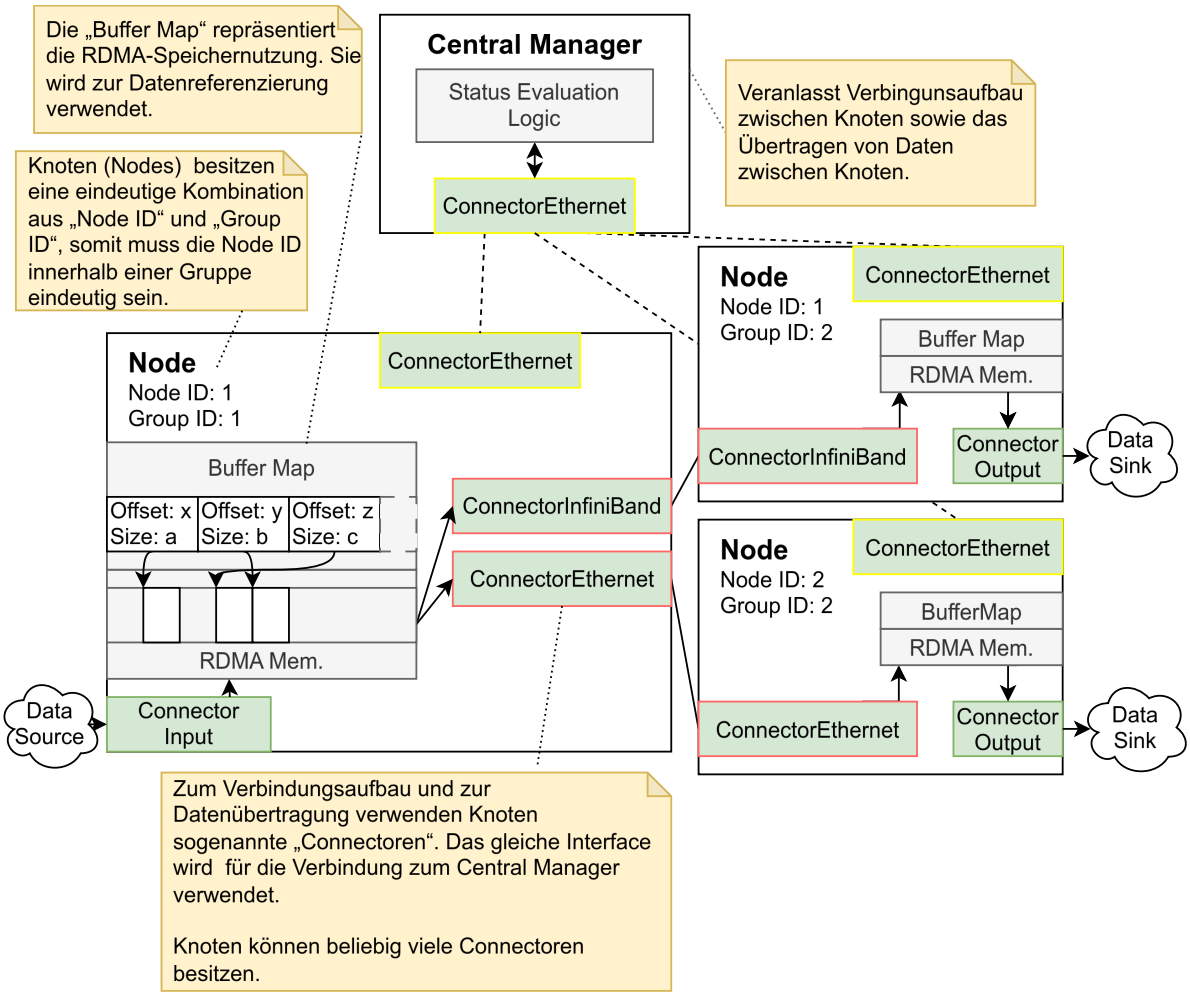

Abbildung 1: Software Architektur mit zentralisiertem Puffer und Übertragungs-Management.

Durch die Unterstützung von Elastizität für Flesnet ist das System flexibler und dynamischer einsetzbar. Zum Beispiel können Analyse-Rechner dynamisch hinzugefügt und entfernt werden, je nach Bandbreiten- und Rechenzeitbedarf. Die praktische Evaluation der entwickelten Lösungen zeigt deren Leistungsfähigkeit.

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.

AP 2: Heterogene Datenvolumina in Flesnet

Das Timeslice Building wird auf dem FLES-Cluster durchgeführt, wobei dauerhaft hohe Bandbreiten zu erreichen sind, um die Datenströme kontinuierlich zusammenzuführen. Der Cluster ist logisch in Eingangs- und Prozessierungsrechner aufgeteilt. Um verteilt Timeslices zu erstellen, benötigt jeder Prozessierungsrechner Daten von allen Eingangsrechnern. Dadurch entsteht eine enge Koppelung zwischen Eingangs- und Prozessierungsrechnern. Um das Netzwerk effizient zu nutzen und Stauungen (Congestion) zu vermeiden, werden die einzelnen Datenübertragungen nach einer rundenbasierten Arbeitsplanung organisiert.

Die Kommunikation zwischen Eingangs- und Prozessierungsrechnern erfolgt mittels Remote Direct Memory Access (RDMA) und entsprechenden Ringpuffern. Sind alle Beiträge zu einer Timeslice in einem Prozessierungsrechner eingetroffen und verarbeitet worden, wird dies den Eingangsrechnern mitgeteilt und der Pufferbereich dort kann erneut genutzt werden. Bei leerem Puffer eines Prozessierungsrechners kann dieser keine Daten verarbeiten. Bei vollem Puffer können die Eingangsrechner keine weiteren Daten an den Prozessierungsrechner schicken. Beide Extreme sollten für einen effizienten Ablauf nach Möglichkeit vermieden werden. Dazu passt eine aktive Datenfluss-Steuerung die Datenrate an und verteilt die einzelnen Datenflüsse zeitlich so, dass Stauungen vermieden werden. Bisher wurde dies primär mit gleich großen Beiträgen je Eingangskanal getestet und untersucht. Im tatsächlichen Experiment muss aber mit heterogenen Größen der einzelnen Beiträge umgegangen werden. Dazu sollten in diesem Arbeitspaket entsprechende Untersuchungen entwickelt, umgesetzt und evaluiert werden.

Rückführung von Ausgangsdaten. Jede Detektor-Aufzeichnung kann einzigartige Anforderungen an die Rechenleistung und Netzwerkgeschwindigkeit aufweisen. Um Flesnet auf diese besonderen Anforderungen testen und auslegen zu können ist es notwendig Detektor-Aufzeichnungen simuliert zu wiederholen. Hierzu wurde das „tsa2msa“ Tool entwickelt, welches aus gebauten Timeslice-Archiven, sogenannten tsa-Dateien, die ursprünglichen Microslice-Datenströme der Eingangskanäle in Form von msa-Dateien rekonstruiert. Diese msa-Dateien können mit dem bereits vorhandenen mstool dem Dateneingang von Flesnet zugeführt werden, um eine echte Detektor-Aufzeichnung wiederzugeben und das Bauen von Timeslices mit Flesnet realitätsnah zu testen.

Für die Untersuchungen haben wir uns Teile der aufgenommenen mCBM-Strahlzeiten von der GSI ins ZIB kopiert, um einen praktischen Eindruck der Heterogenität der Datenvolumina im CBM-Projekt zu erhalten. Für die Analyse haben wir uns Daten aus zwei vom CBM Projekt als „Golden-Runs“ bezeichneten Strahlzeiten, also als qualitativ gut und ausreichend repräsentativ eingeschätzt, kopiert. Dies umfasst den Lauf Nummer 2391 mit Nickel auf Nickel Kollisionen vom 26. Mai 2022 (57 GB Datenvolumen) und den Lauf Nummer 2488 mit Gold auf Gold Kollisionen vom 18. Juni 2022 (511 GB von insgesamt 9.5 TB) jeweils mit den Subsystemen TRD, STS, TRD2D, MUCH, TOF, BMON und RICH. (insgesamt 560 GB). Anschließend konnten die Microslice-Datenströme erfolgreich mit dem entwickelten „tsa2msa“-Tool wieder hergestellt werden.

Wiedergabe mit heterogenen Datenvolumina. Für die Unterstützung und Evaluation von Flesnet mit heterogenen Datenvolumina haben wir sowohl die rückgeführten Daten aus echten mini-CBM-Strahlzeiten als Grundlage für eine realistische Wiedergabe auf der ZIB-Infrastruktur wiedergegeben, als auch Flesnet mit synthetisch erzeugten Daten und Datenverteilungen evaluiert. Kritisch ist bei der Wiedergabe von echten Messdatenströmen das Datenvolumen, das vom persistenten Speicher gelesen werden muss und die dabei erreichbare Geschwindigkeit. In größerem Maßstab ist es nicht realistisch die von CBM anvisierten Quell-Datenraten auf diese Art simulieren zu können. Hierfür reichen aber grundsätzlich die Metadaten der tatsächlichen mCBM-Läufe aus, um entsprechende Datenpakete für eine Datenwiedergabe zu erzeugen, mit einer Größenverteilung über die Kanäle und die Zeit, die den tatsächlichen Lauf nachbildet. Hierfür haben wir entsprechend die Metadaten aus den Strahlzeiten extrahiert und separat gespeichert. Diese Daten dienen dann der Steuerung eines größenkonformen Replays. Als tatsächliche Daten werden dabei vorgenerierte Zufallsdaten aus einem großen sharedmemory Puffer versendet. Mit diesem Ansatz lassen sich realistische Datenverteilungen und hohe

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.

Bandbreiten für Testzwecke realisieren. Zusätzlich lässt sich die Abspielgeschwindigkeit auch je nach gewünschter Datenrate im Experiment im Vergleich zur tatsächlichen Aufzeichnung beschleunigen oder verlangsamen.

Durch entsprechende Experimente wurde das Verhalten von Flesnet mit praxisähnlichen heterogenen Datenvolumina untersucht und entsprechende Anpassungen im Hinblick auf die Lastverteilung und das Datenstrom- und Puffermanagement entwickelt, um dauerhaft hohe aggregierte Bandbreiten beim Bauen von Timeslices zu erreichen.

AP 3: Skalierung, mCBM-Integration

Die Skalierbarkeit von Flesnet wurde weiter untersucht und auch im Hinblick auf Überlastsituationen verbessert. Dies zeigen zum Beispiel unsere Ergebnisse, die auch in den ersten technische Entwurfsbericht „Technical Design Report for the CBM – Online Systems – Part I“ des CBM-Projekts für zentrale Systeme eingeflossen sind. Dieser Entwurf wurde entsprechend der Empfehlung des Expertenausschusses Experimente (ECE) nach fachlicher Begutachtung von FAIR im Juli 2023 genehmigt und veröffentlicht [2] (DOI: 10.15120/GSI-2023-00739). Der Bericht beschreibt die Hardware-Architektur des

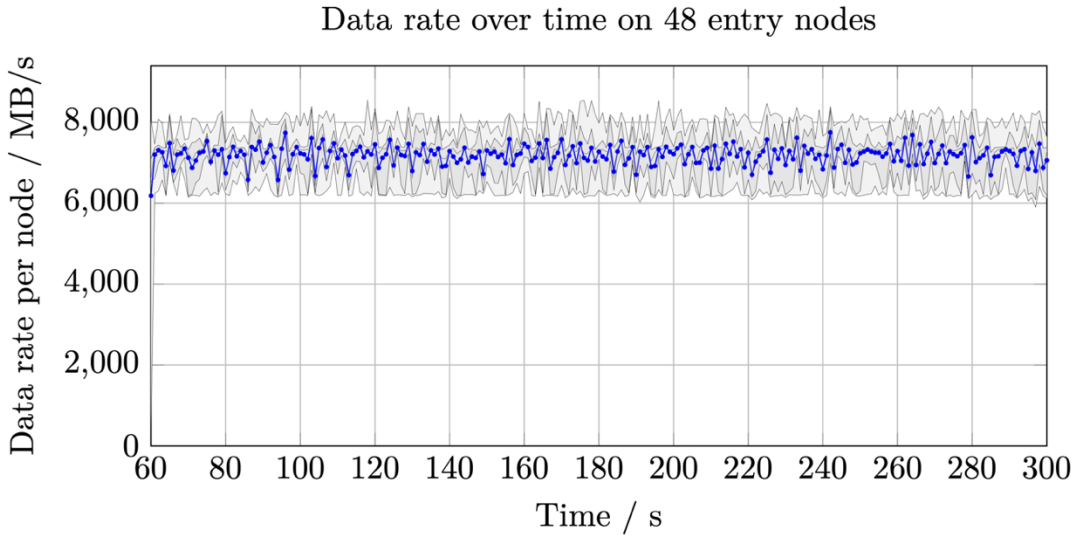

Abbildung 1: Gemessene durchschnittliche Datenrate zum Bauen von Timeslices je Eingangsknoten. Graue Bereiche und Linien geben Minimum/Maximum, die 10./90. Perzentile sowie den Median an [2].

FLES-Entry-Clusters, das Netzwerk zur Erstellung von Zeitscheiben (einschließlich Flesnet) und den Datenpfad bis zur Erstellung der Zeitscheiben. Zu dem Bericht haben wir unter anderem unsere Erfahrungen und Messungen zur Skalierbarkeit und dauerhaften Leistungsfähigkeit von Flesnet beigetragen. Mit insgesamt 96 Rechnern (48 Eingangsrechner und 48 Prozessierungsknoten) konnte stabil eine aggregierte Datenrate von bis zu 340 GB/s erreicht werden (siehe Abbildung 1).

Im mCBM-Projekt werden alle wichtigen Komponenten des CBM-Projektes in kleinem Maßstab integriert und gemeinsam eingesetzt. In sogenannten „Dry-Runs“, an denen wir uns auch beteiligt haben, wird die Prozessierungskette getestet – entweder auf der Basis früherer Datenaufzeichnungen, oder auf der Basis des tatsächlichen mCBM-Experimentaufbaus ohne expliziten Teilchenstrahl, sondern auf Basis der kosmischen Strahlung (Cosmic-Run). Die Dry-Runs bieten gute Gelegenheiten für Integrationstests und zum Sammeln praktischer Erfahrungen, typischer Fehlersituationen und -ursachen sowie deren Lokalisation und Behebung. So konnte Flesnet weiter erprobt werden und anhand der auftretenden Ausfälle und Probleme im alltäglichen Praxisbetrieb weiter verbessert werden.

Für die Flesnet-Entwicklung am ZIB können wir mittlerweile auch auf eine Infiniband basierte Clusterinfrastruktur zurückgreifen. Hier stehen uns insgesamt ca. 100 Knoten mit Infiniband 100 Gbps (2x IB-

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.

HDR) Datenrate zur Verfügung. Um diese für unsere lokalen Experimente zu nutzen, wurden entsprechende Anpassungen an den Job-Scripten, aber auch an der Clusterkonfiguration selbst in Absprache mit den Systemadministratoren durchgeführt.

Zur Code-Pflege und Modernisierung wurden kleinere Beiträge im laufenden Entwicklungsprozess geleistet. Zum Beispiel werden einige veraltete Funktionen jetzt explizit als „deprecated“ gekennzeichnet, wodurch eine Warnung ausgegeben wird, wenn die Funktion weiterhin in Client-Code wie CBMroot verwendet wird. Der Code wurde an neuere Compiler- und Bibliotheksversionen angepasst, was für einige Code-Konstrukte neue Warnungen auslöste. In ähnlicher Weise wurde die minimal erforderliche CMake- und Bibliotheksversion erhöht und an die aktuell verwendeten Linux-Distributionen in CBM angepasst [7].

Wir haben die Dokumentation in verschiedenen Bereichen erweitert und aktualisiert, um neue Nutzer und Entwickler besser zu unterstützen. Die Projekteinstellungen für Doxygen, einem Tool zum Extrahieren und Präsentieren von C++-Code-Dokumentation, wurden für eine ordnungsgemäße Funktion aktualisiert und um einige zusätzliche Dokumente ergänzt, die wir separat bereitstellen. Wir stellen ein Tutorial für eine Flesnet-Einrichtung mit InfluxDB- und Grafana-Überwachung in Docker-Containern auf einer lokalen Maschine bereit. Eine solche Einrichtung unterstützt die Entwicklung detaillierterer Überwachungsmöglichkeiten mit lokalen Tests, bevor diese schließlich freigegeben werden. Für einen beispielhaften minimalen Flesnet-Lauf haben wir eine Anleitung zum Einrichten einer Datenverarbeitungskette in Software verfasst. Sie (1) generiert Microslices und schreibt sie in den gemeinsam genutzten Speicher. (2) verwendet einen Flesnet-Eingangsprozess, um die Microslices aus dem gemeinsam genutzten Speicher zu lesen, Timeslices zu erstellen und sie in einen anderen gemeinsam genutzten Speicherbereich zu schreiben; und (3) liest die erstellten Timeslices und schreibt sie in einem abschließenden Verarbeitungsschritt in ein Timeslice-Archiv. Wir beschreiben die relevanten Kommandozeilenparameter und ihre Abhängigkeiten und erklären, wie sie verwendet werden [7].

Kooperation mit Dritten und sonstige Aktivitäten

Zur Abstimmung im Verbund haben wir regelmäßig an den CBM Collaboration Meetings an der GSI/FAIR in Darmstadt und anderen Treffen kleinerer Gruppen teilgenommen, unsere Arbeiten dort vorgestellt und diskutiert. Außerdem haben wir an den regelmäßigen stattfindenden Collaboration Board Meetings teilgenommen.

Außerdem haben wir die engere Zusammenarbeit bei der Flesnet-Entwicklung in regelmäßigen Videokonferenzen mit der AG Prof. Lindenstruth am FIAS abgestimmt.

C.B.M. Retreat Limburg (27. – 28.02.2024): Die Arbeitsteilung und Zusammenarbeit des gesamten Verbundes bündelt sich im Verbund „ErUM FSP T06 C.B.M.“, der die Beteiligung der deutschen Partner am internationalen CBM-Projekt darstellt. Zur Abstimmung und strategischen Planung haben wir den vom Lehrstuhl von Prof. Joachim Stroth organisierten C.B.M. Retreat in Limburg besucht, bei dem auch Vertreter des Projektträgers teilgenommen haben. Der direkte Austausch und Kontakt in diesem Rahmen, hat die Zusammenarbeit weiter positiv unterstützt.

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

Für das Erreichen der Projektziele und die in diesem Zusammenhang notwendigen Forschungs- und Entwicklungsarbeiten wurde geeignetes, qualifiziertes Personal für die Projektarbeit entsprechend dem genehmigten Projektantrag eingesetzt und vergütet. Im Zusammenhang mit dem Projekt stehende notwendige Dienstreisen, Forschungsaufenthalte und Nebenkosten wurden finanziell unterstützt. Details sind dem zahlenmäßigen Nachweis zu entnehmen. Außerdem wurden Mittel für den CBM Construction Common Fund bewilligt und entsprechend eingezahlt.

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.

5 Erzielte Ergebnisse mit Gegenüberstellung der vereinbarten Ziele

Folgende Meilensteine waren geplant und sind zum Teil mit Verzögerung und mit Hilfe einer kostenneutralen Verlängerung der Projektlaufzeit erreicht worden.

AP1: Elastizität in Flesnet

A1, Q2-2023: Entwicklung und Implementierung der Elastizität für Knoten in Flesnet ist weitestgehend abgeschlossen (AP1)

Der Meilenstein war wegen Personalengpässen deutlich verzögert und wurde erst im Frühjahr 2024 erfolgreich erreicht. Die Elastizität bei Ausfällen oder dem dynamischen Hinzufügen von Knoten konnte erfolgreich demonstriert werden.

A2, Q1-2024: Evaluation der Elastizität für Knoten in Flesnet wurde durchgeführt (AP1)

Der Meilenstein konnte zum Ende der kostenneutral verlängerten Projektlaufzeit erfolgreich erreicht werden. Die Elastizität von Knoten wurde erfolgreich demonstriert.

AP 2: Heterogene Datenvolumina in Flesnet

B1, Q2-2023: Entwicklung und Implementierung für heterogene Datenvolumina in Flesnet ist weitestgehend abgeschlossen (AP2)

Der Meilenstein war wegen Personalengpässen deutlich verzögert und wurde erst im Frühjahr 2024 erfolgreich erreicht. Teile der Entwicklung waren, wie oben beschrieben, Tools für ein Replay mit heterogenen synthetischen und aus dem mini-CBM während Strahlzeiten gesammelten Daten, um Flesnet entsprechend für den Umgang mit heterogenen Datenvolumina anzupassen.

B2, Q1-2024: Evaluation der Unterstützung für heterogene Datenvolumina in Flesnet wurde durchgeführt (AP2)

Der Meilenstein konnte zum Ende der kostenneutral verlängerten Projektlaufzeit erfolgreich erreicht werden. Teile der Entwicklung waren, wie oben beschrieben, die Evaluation auf der Basis von echten mini-CBM Strahlzeitdaten, die für ein realistisches Replay genutzt wurden.

AP 3: Skalierung, mCBM Integration

C1, Q1-2024: Die Skalierbarkeit von Flesnet wurde weiter verbessert und evaluiert (AP3)

Der Meilenstein war wegen Personalengpässen verzögert. Er konnte zum Ende der kostenneutralen Verlängerung des Projekts erfolgreich erreicht werden. Ergebnisse bezüglich der Skalierbarkeit sind auch in den „*Technical Design Report for the CBM Online Systems – Part I, DAQ and FLES Entry Stage“* [2] eingeflossen.

C2, Q1-2024: Die an Flesnet weiter- und neuentwickelten Funktionalitäten wurden in mini-CBM integriert und evaluiert (AP3)

Der Meilenstein war wegen Personalengpässen verzögert. Er konnte zum Ende der kostenneutralen Verlängerung des Projekts erfolgreich erreicht werden. Die im Projekt entwickelten Erweiterungen und Verbesserungen wurden auch auf der Infrastruktur des mini-CBM erfolgreich getestet und evaluiert.

6 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Das CBM-Experiment am GSI/Fair benötigt Flesnet als wichtige Komponente auf dem Datenverarbeitsungspfad. Flesnet aggregiert und leitet die Messdatenströme in Zeitscheiben zu Analyserechnern. Die

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.

im Projekt adressierten Ziele der Unterstützung von Elastizität und heterogenen Datenvolumina für Flesnet sind dabei im geplanten praktischen Einsatz des Systems wichtige Eigenschaften. Durch die entwickelten Anpassungen bezüglich dieser Aspekte im Rahmen dieses Projekts konnte die Unterstützung des Systems für Elastizität und heterogene Datenvolumina deutlich verbessert werde, was auch unsere praktische Evaluation zeigt.

7 Voraussichtlicher Nutzen, insbesondere Verwertbarkeit der Ergebnisse

Flesnet ist für das CBM-Experiment am FAIR integraler Bestandteil der Datenverarbeitungskette als Bindeglied zwischen dem Detektor und der inhaltlichen physikalischen Analyse der beobachteten Teilchenkollisionen. Eine effiziente, skalierbare, und ausfalltolerante Messdatenverarbeitung und das Erstellen der Zeitscheiben mit Flesnet sind für die Projektziele von CBM unumgänglich. Die durchgeführten Projektarbeiten tragen in den Bereichen Elastizität und heterogene Datenvolumina wesentliche Erweiterungen und Verbesserungen dazu bei.

Die entwickelte Software und Flesnet selbst stehen als OpenSource zur Verfügung (https://github.com/cbm-fles/flesnet) und können in der CBM-Collaboration und von Dritten darüber hinaus genutzt und weiterentwickelt werden.

Der tatsächliche Aufbau des CBM-Experiments und des Teilchenbeschleuniger SIS100 schreiten stetig voran, wie auch ein Blick auf die Baustelle am GSI/FAIR verrät. Für den tatsächlichen Einsatz im Experimentbetrieb sind weitere Untersuchungen, Stabilisierungen und Anpassungen an neue Gegebenheiten von Flesnet notwendig, die in einer weiteren Projektförderphase adressiert werden sollten.

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

Die Kommunikationsbibliothek „libfabric“ auf der unsere Arbeiten basieren wird kontinuierlich weiterentwickelt. Im Juli 2021 – quasi zum Beginn dieses Projektes – wurde Version 1.13.0 veröffentlicht. Im Juli 2024 ist dann Version 1.22.0 veröffentlicht worden und zum Dezember 2024 erfolgte der Switch auf Version 2.0.0. Über die Projektlaufzeit hinweg sind insgesamt neun und inklusive kleinerer Versionssprünge 25 Versionen allein dieser Bibliothek erschienen. Für uns heißt dies im Entwicklungsprozess diese jeweils entsprechend zu berücksichtigen indem wir die Kompatibilität unserer Software mit den neuen Änderungen sowohl im Hinblick auf die Grundfunktionalität als auch die Leistungsparameter überprüfen und gegebenenfalls Flesnet entsprechend anpassen.

9 Erfolgte und geplante Veröffentlichungen der Ergebnisse

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

keine

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

Farouk Salem, Florian Schintke. *Large-Scale Performance of the Data-Flow Scheduler (DFS) and FLESnet.* CBM Progress Report 2021, pp. 170-171, 2022, doi: 10.15120/GSI-2022-00599.

CBM Collaboration. *Technical Design Report for the CBM Online Systems – Part I, DAQ and FLES Entry Stage.* J. d. Cuveland, D. Emschermann, V. Friese, I. Fröhlich, P. Gasik, D. Hutter, W. Müller, and C. Sturm (editors). Technical Report, Darmstadt, 2023. FAIR Technical Design Report. URL: https://repository.gsi.de/record/340597, doi:10.15120/GSI-2023-00739.

Florian Schintke, Nico Greve, Jan de Cuveland. *Recent Flesnet developments and work in progress.* CBM Progress Report 2023, p. 132, 2024, doi: 10.15120/GSI-2024-00765.

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.

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

keine

Literatur

[1] P. Buncic, M. Krzewicki, and P. Vande Vyvre. *Technical Design Report for the Upgrade of the Online-Offline Computing System.* Technical Report CERN-LHCC-2015-006. ALICE-TDR-019, June 2015. URL: https://cds. cern.ch/record/2011297.

[2] CBM Collaboration. *Technical Design Report for the CBM Online Systems – Part I, DAQ and FLES Entry Stage.* J. d. Cuveland, D. Emschermann, V. Friese, I. Fröhlich, P. Gasik, D. Hutter, W. Müller, and C. Sturm (editors). Technical Report, Darmstadt, 2023. FAIR Technical Design Report. URL: https://repository.gsi.de/record/340597, doi:10.15120/GSI-2023-00739.

[3] S. Chunduri, T. Groves, et al. *GPCNeT: designing a benchmark suite for inducing and measuring contention in HPC networks.* In M. Taufer, P. Balaji, and A. J. Peña, editors, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2019, Denver, Colorado, USA, November 17-19, 2019*, pages 42:1–42:33. ACM, 2019. doi:10.1145/3295500.3356215.

[4] D. De Sensi, S. Di Girolamo, K. H. McMahon, D. Roweth, and T. Hoefler. *An in-depth analysis of the slingshot interconnect. CoRR*, abs/2008.08886, 2020. URL: https://arxiv.org/abs/2008.08886, arXiv:2008.08886.

[5] F. Salem, F. Schintke. *Large-Scale Performance of the Data-Flow Scheduler (DFS) and FLESnet.* CBM Progress Report 2021, pp. 170-171, 2022, doi: 10.15120/GSI-2022-00599.

[6] F. Salem, F. Schintke, T. Schütt, and A. Reinefeld. *Scheduling data streams for low latency and high throughput on a Cray XC40 using Libfabric.* Concurr. Comput. Pract. Exp., 32(20), 2020. doi:10.1002/cpe.5563.

[7] F. Schintke, N. Greve, J. de Cuveland. *Recent Flesnet developments and work in progress.* CBM Progress Report 2023, p. 132, 2024, doi: 10.15120/GSI-2024-00765.

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.