Sachbericht Verwendungsnachweis (Teil 1) - 16ME0693	
Zuwendungsempfänger	ParTec AG
Verbundprojekt	**IFCES2** – Optimierung von Simulationsalgorithmen für Exascale-Supercomputer zur Berechnung des Erdsystemmodells ICON
Förderkennzeichen	16ME0693
Bewilligungszeitraum	01.10.2022 – 31.12.2025

Teil I: Kurzbericht

Das IFCES2-Projekt zielte darauf ab, die Skalierbarkeit des Erdsystemmodells ICON auf heterogenen Exascale-Supercomputing-Systemen zu verbessern. Dazu sollten neue Methoden für Parallelisierung, Kommunikation und dynamische Lastverteilung entwickelt und Optimierungen insbesondere hinsichtlich funktionaler Nebenläufigkeit und des damit verbundenen Datenaustauschs in einem Co-Design-Ansatz durchgeführt werden. Dies diente dem übergeordneten Ziel, die Unsicherheiten in Klimaprojektionen zu verringern und die Klimaforschung durch neue Möglichkeiten der hochauflösenden Modellierung zu stärken.

Ursprüngliche Aufgabenstellung

ParTec war einer von sechs Verbundpartnern im Projekt, das sich wiederum in insgesamt sechs Arbeitspakete (AP) gliederte. Dabei spielte ParTec eine zentrale Rolle in AP3 (*Modulares Supercomputing und MPI-Tuning*), dessen Leitung ParTec sich mit dem FZJ teilte. Die ursprüngliche Aufgabenstellung für ParTec umfasste dabei vor allem die folgenden Arbeitspunkte, in denen sich ParTec auf die Optimierung von ParaStation MPI hinsichtlich der Nutzung durch ICON fokussierte:

Analyse und Anpassung von Übertragungsprotokollen. Zentraler Aspekt waren hier die verschiedenen Kommunikationsprotokolle von ParaStation MPI. Diese sollten in ihrer Verwendung durch ICON genauer analysiert und daraufhin speziell für ICON angepasst werden, um die Übertragungsleistung zu optimieren. Neben den eigentlichen Protokollen standen hier deren individuelle Parameter sowie deren Übergangsschwellwerte untereinander im Fokus der Analysen.

Gezielte Optimierung leistungskritischer Codepfade. Ziel dieser Teilaufgabe war es, Engpässe in den leistungskritischen Codeteilen und Aufrufpfaden innerhalb von ParaStation MPI bei seiner Nutzung durch ICON zu identifizieren, zu analysieren und durch gezielte Anpassungen für ICON zu optimieren. Dieser Arbeitspunkt adressierte damit insbesondere mögliche Verbesserungen an der inneren Ereignisschleife von ParaStation MPI bzw. seiner tieferschichtigen Kommunikationsbibliothek pscom.

Optimierte Behandlung von abgeleiteten MPI-Datentypen. Der Schwerpunkt dieser Teilaufgabe adressierte wiederum die höheren Schichten von ParaStation MPI, in denen zum Beispiel neben der Umsetzung von kollektiven Kommunikationsmustern auch die Umsetzung von abgeleiteten Datentypen angesiedelt ist. Da solche Datentypen eine wichtige Rolle für ICON und seine Kopplungskomponente YAXT spielen, zielte diese Teilaufgabe auf mögliche Optimierungen in diesen Schichten von ParaStation MPI ab.

Das Hauptziel des Teilvorhabens von ParTec innerhalb des IFCES2-Projekts lässt sich somit insgesamt als die Analyse und die Umsetzung von leistungsoptimierenden Anpassungen von ParaStation MPI an die Bedürfnisse von ICON zusammenfassen.

Ablauf des Vorhabens

In den Anfangsphasen des Projekts konzentrierte sich ParTec auf die Einarbeitung im Umgang mit ICON, seinen Komponenten sowie seinen typischen Test- und Anwendungsfällen. Da zu dieser Zeit ICON noch nicht als Open-Source-Software zur Verfügung stand, wurden die dabei gemachten Leistungsuntersuchungen zunächst als *Black-Box*-Analysen durchgeführt. In diese initiale Einarbeitungsphase fielen neben dem Kick-off-Treffen im März 2023 in Hamburg auch diverse Online-Meetings und technische Arbeitstreffen, in denen Informationen, Hinweise und Hilfestellungen zwischen den MPI-Entwicklern von ParTec und den ICON-Experten der anderen Verbundpartner ausgetauscht wurden.

Mit der Öffnung von ICON als frei zugängliches Open-Source-Produkt konnten ab Anfang 2024 die bisher gewonnenen Erkenntnisse durch Codeanalysen verifiziert und erweitert werden. Dabei stellten sich die bis dahin durchgeführten allgemeinen Maßnahmen zur Leistungssteigerung als weniger zielführend heraus als zunächst angenommen. Basierend auf den gewonnenen Kommunikationsprofilen und den daraus abgeleiteten Anforderungen wurden in dieser Projektphase dann gezielt ICON-spezifische Analysen und entsprechende Anpassungen des leistungskritischen Codepfades der pscom-Bibliothek durchgeführt.

Als Ergebnis lag mit Erreichen des Meilensteins MS2.3 und damit zum entsprechenden Projekttreffen im November 2024 in Leipzig eine bereits stark auf die Bedürfnisse von ICON angepasste Version von ParaStation MPI vor. In den anschließenden Phasen des Projekts wurden zum einen diese Anpassungen weiter durch ParTec verfeinert, zum anderen wurde aber auch mit den Arbeiten an einer Verallgemeinerung der Optimierungen und einer Integration bestimmter Anpassungen in den Hauptentwicklungszweig begonnen, um diese Ergebnisse nicht nur ICON, sondern einer größeren Klasse an Anwendungen auf Basis von ParaStation MPI verfügbar zu machen. Diesen Arbeiten kam dabei auch die bewilligte kostenneutrale Verlängerung des Projekts um drei Monate zugute.

Zum Ende des Projekts und mit Erreichen des Meilensteins MS3.2 konnte ParTec so am Abschlusstreffen im Januar 2026 in Hamburg die im Kontext der entwickelten Optimierungen und prototypischen Erweiterungen erzielten Leistungssteigerungen abschließend präsentieren und gemeinsam mit den Verbundpartnern die erzielten Fortschritte sowie den zukünftigen Transfer der Ergebnisse diskutieren.

Wesentliche Ergebnisse

Aus Sicht der ParTec AG ist das wesentliche Ergebnis des IFCES2-Projekts eine grundlegend überarbeitete Version der inneren Ereignisschleife (sog. *Progress-Engine*) der pscom-Schicht von ParaStation MPI. Durch die auf ICON zugeschnittenen Veränderungen, die im Rahmen von IFCES2 hier vorgenommen wurden, konnten die Verweilzeiten in diesen Codeabschnitten deutlich verkürzt werden, was sich in einer gemessenen Reduzierung von bis zu 25% der durchschnittlichen Wartezeit und von bis zu 10% der minimalen Wartezeit manifestiert. Zudem zeigten Analysen des zentralen Kommunikationsschemas von ICON vor den Optimierungen große Unterschiede im Laufzeitverhalten zwischen den Prozessen, die durch die erfolgten Anpassungen deutlich verringert werden konnten. Somit kann bei der Verwendung dieser speziell zugeschnittenen Version von ParaStation MPI auch von einer gleichmäßigeren und somit effizienteren Nutzung nicht nur der Kommunikations-, sondern auch der Rechenressourcen ausgegangen werden. Und obwohl diese entwickelten Optimierungen in erster Linie auf ICON ausgerichtet waren, konnten auch in synthetischen Benchmarks deutliche Leistungsgewinne – insbesondere hinsichtlich der Kommunikationslatenzen – beobachtet werden. Diese Ergebnisse deuten damit darauf hin, dass sich die Ansätze der entwickelten Optimierungen sowie die dabei gewonnene Expertise zukünftig auch auf andere Anwendungen übertragen lassen.