

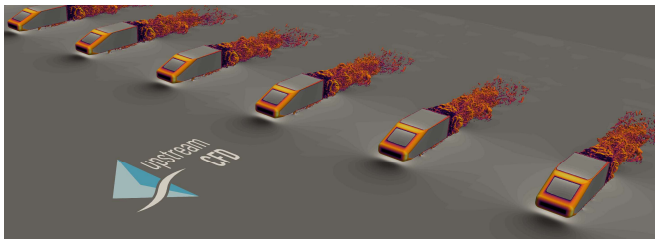
## Abschlussbericht: Teil 1 - Kurzbericht

### Ausgangssituation und wissenschaftlich-technischer Stand

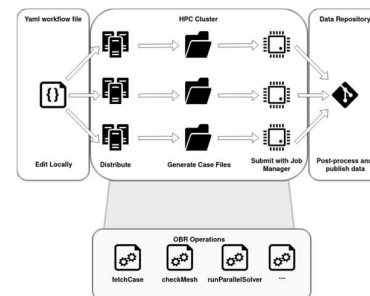
Im Projekt EXASIM („Exascale-fähige Softwarewerkzeuge zur Strömungssimulation im industriellen Design- und Optimierungsprozess“) wurde untersucht, wie sich moderne GPU-basierte Hochleistungsrechner, effizient für industrielle Strömungssimulationen einsetzen lassen. Der Ausgangspunkt des Vorhabens war die ursprünglich für CPUs entwickelte Open-Source-Software OpenFOAM, welche über eine vom Projektpartner TUM entwickelte Schnittstelle (OpenFOAM Ginkgo Layer - OGL) mit der für GPUs-optimierte lineare Algebra Bibliothek Ginkgo gekoppelt werden sollte. Dabei wurde der hybride Ansatz verfolgt, bei dem nur die Lösung der linearen Gleichungssysteme von Druck- und Impulsgleichungen auf die GPU ausgelagert wird und die restlichen Teile des Lösungsalgorithmus auf der CPU verbleiben. Vorstudien an wissenschaftlichen Testfällen versprachen bereits Effizienzsteigerungen, ohne deutlich aufwendigere Ansätze wie die Portierung des gesamten Algorithmus auf GPU oder ein vollständiges Re-Design eines GPU-nativen Strömungslösers durchführen zu müssen. Die Rolle der Upstream CFD GmbH (UCFD) war dabei die Praxistauglichkeit der entwickelten Methoden an realistischen, industriellen Anwendungsfällen zu überprüfen und zu bewerten.

### Ablauf des Vorhabens

Ein zentraler Bestandteil der Arbeiten war der Aufbau einer reproduzierbaren und automatisierten Benchmark-Umgebung. Dazu wurde eine umfangreiche Suite von Testfällen aus verschiedenen Bereichen der CFD zusammengestellt, darunter simple Validierungsfälle wie Kanalströmungen sowie komplexe industrielle Anwendungen aus der Fahrzeug- und Umweltströmung, siehe Abbildung 1a. Jeder Testfall wurde mit einem automatisierten Workflow versehen, der Gittererstellung, Simulation und Auswertung umfasst.



a) Strömungsvisualisierung des für Skalierungstests 6-fach verketteten WindsorBody-Falls



b) Schematische Darstellung des OBR Workflows

**Abbildung 1: Darstellungen zum HPC-Workflow und der Testfallsuite.**

Zur Orchestrierung dieser Studien wurde das Softwarewerkzeug OpenFOAM Benchmark Runner (OBR) entwickelt und eingesetzt. Dieses Python-basierte Framework ermöglicht es, Parameterstudien systematisch aufzusetzen, Simulationen automatisiert auf HPC-Systemen auszuführen und die Ergebnisse strukturiert zu sammeln und auszuwerten, siehe Abbildung 1b.

Im nächsten Schritt wurden umfangreiche Benchmarkstudien durchgeführt. Dabei wurde untersucht, in welchem Maße der hybride Ansatz eine Beschleunigung für realistische industrielle Simulationen

ermöglicht. Eine wichtige Komponente war die Definition geeigneter Metriken, da der CPU-GPU-Vergleich durch die unterschiedlichen optimalen Skalierungspunkte nicht trivial ist. (~20 000 Zellen/CPU-Kern vs. Millionen Zellen/GPU). Neben dem Vergleich des Fall- und Löser-spezifischen Durchsatzes (gemessen in Finite-Volumen Operationen pro Sekunde – FVOPS) und der Effizienz (FVOPS pro Recheneinheit) wurden nutzerorientierte Größen wie Zeit-, Energie- und Kosten-zu-Lösung berechnet, um die Wirtschaftlichkeit direkt bewerten zu können.

## Wesentliche Ergebnisse des Projekts

Die durchgeführten Untersuchungen zeigen, dass der hybride Ansatz (nur Auslagerung des linearen Löser auf GPU) für realistische Anwendungen nur in Spezialfällen Effizienzgewinne liefert. Für die meisten der 8 untersuchten Fälle der Testfallsuite zeigte sich, dass ein erheblicher Anteil der Rechenzeit auf die auf der CPU verbleibende Assemblierung der Matrizen entfällt und damit die erreichbare Gesamtbeschleunigung begrenzt. Ein wesentliches strategisches Ergebnis des Projekts ist daher die Erkenntnis, dass langfristig ein vollständig GPU-nativer CFD-Solver erforderlich ist, um das Potenzial moderner Beschleunigerarchitekturen auszuschöpfen, siehe Abbildung 2. SPUMA<sup>1</sup>, ein von CINECA entwickelter OpenFOAM-GPU-Port, zeigt bereits eine deutlich höhere Performance und erreicht in Tests eine Kostenreduktion bis zu einem Faktor von 3.5 sowie eine Energieeinsparung von bis zu einem Faktor 2 gegenüber CPU-basierten Simulationen.

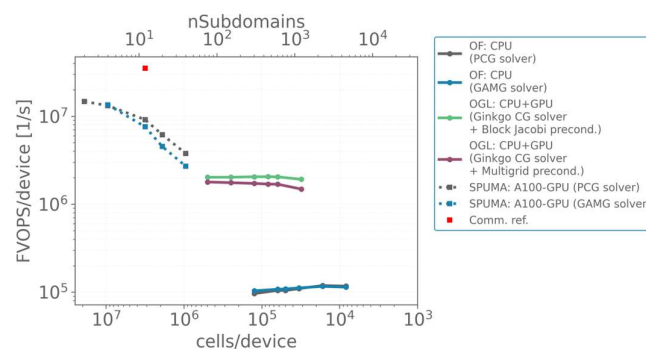


Abbildung 2: Vergleich des Durchsatzes pro Recheneinheit WindsorBody6x Fall.

Vergleiche mit kommerziellen GPU-Lösern indizieren ein nochmal 3-fach höheres Performance-Potential. Ein wesentlicher Grund hierfür liegt in dem altersbedingten technischen Zustand der OpenFOAM-Codebasis, deren Struktur und Datenlayouts bei einer reinen Portierung auf GPU weiterhin strukturelle Effizienzgrenzen verursachen. Daher wurde gemeinsam mit der TU München die Entwicklung eines neuen GPU-nativen Solver-Backends unter dem Namen NeoN<sup>2</sup> begonnen. Erste Implementierungen ohne weitergehende Optimierungen erreichen bereits eine vergleichbare Performanz wie der portierte SPUMA-Ansatz.

Mit dem Framework aus Testfall-Suite und OBR kann nun systematisch die weitere Entwicklung der verschiedenen GPU-Solver-Ansätze mit automatischen Parameterstudien begleitet werden, um deren Performance-Entwicklung kontinuierlich zu analysieren und zu bewerten. Zusätzlich bietet sich auch die Untersuchung des Einflusses neuer Hardware-Generationen an.

<sup>1</sup> <https://gitlab.hpc.cineca.it/exafoam/spuma>

<sup>2</sup> <https://github.com/exasim-project/NeoN>