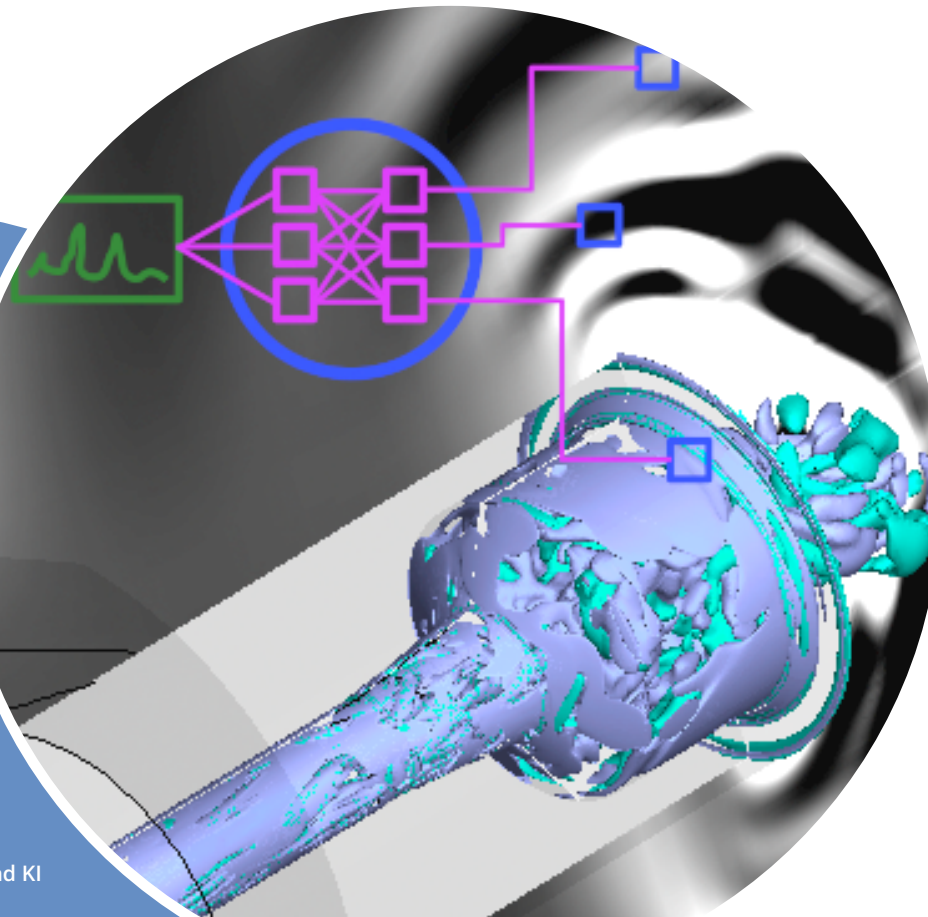




Effiziente Simulationen durch intelligente Parallelisierung.

ResCHKI hat es uns im Projekt HPC4nuberisim2AI ermöglicht, HPC und KI so zu verbinden, dass komplexe Simulationen effizient und ressourcenschonend durchgeführt werden können.



Erfolgsgeschichte

Das Projekt „ResCHKI“ unterstützt vor allem KMU dabei, ihre Prozesse, Produkte und Dienstleistungen mithilfe moderner Daten- und Rechenverfahren wie High-Performance-Computing (HPC), Data Analytics und KI effizienter und nachhaltiger zu gestalten. Durch Beratung, Vernetzung, Veranstaltungen und praxisnahe Umsetzungen werden technische Hürden abgebaut und digitale Lösungen erprobt.

Das Projekt wird vom Ministerium für Umwelt, Klima und Energiewirtschaft Baden-Württemberg gefördert.

Nuberisim ist ein in Karlsruhe ansässiges Ingenieurbüro, das sich auf numerische Strömungssimulationen und aeroakustische Analysen spezialisiert hat. Das Unternehmen unterstützt industrielle Kunden bei der Entwicklung, Optimierung und Lärminderung strömungsführender Komponenten wie Pumpen, Lüfter oder Turbinen. Mithilfe hochauflösender Simulationen liefert Nuberisim fundierte Entscheidungsgrundlagen, um Leistungsfähigkeit, Effizienz und Akustik technischer Systeme gezielt zu verbessern.

Das Pilotprojekt

Im Pilotprojekt „**HPC4nuberisim2AI**“ wurde ein innovativer Ansatz verfolgt, um CFD- und aeroakustische Simulationen ressourceneffizienter durchzuführen. Mit Hilfe eines KI-basierten Surrogatmodells, das einen Teil der Strömungssimulation (CFD) und der aeroakustischen Prognose ersetzt, konnten bereits Einsparungen an der Gesamtlaufzeit eines Simulationsprozesses erzielt werden. Die Einschränkung des KI-Modells auf nur eine GPU führt jedoch zu Verlusten in Bezug auf die erzielbare Ergebnismenge der Simulation, was für realistische Anwendungsfälle nicht nur inakzeptabel, sondern auch ineffizient bezüglich Rechenzeit und Speicherbedarf ist. Ziel ist daher, das KI-Modell, insbesondere Graph-Neural-Networks, effizient zu parallelisieren, wodurch eine Optimierung von Ressourcenbedarf (Rechenleistung, Speicherbedarf) und Prognosegüte auf High-Performance-Computing-Clustern realisiert werden kann. Dadurch zeigt das Projekt exemplarisch, wie die intelligente Verbindung von KI und HPC die Wirtschaftlichkeit und Qualität transienter Simulationen steigert.

Die Herausforderung

Für das bestehende KI-Modell muss jeder 3D-CFD-Ergebnisdatensatz der Zeitreihe des gewählten Anwendungsfalls von durchschnittlicher Größe signifikant beschnitten werden. Dadurch wird die Simulation

insgesamt ineffizient, da ein großer Teil der CFD-Ergebnisse schlicht verworfen werden muss, und so die Rechenzeiterparnis mit Hilfe des KI-Modells durch Verluste an der Prognosegüte teuer erkauft wird. Darüber hinaus sind manuelle Eingriffe erforderlich, die mit Blick auf andere Anwendungsfälle eine effiziente Automatisierung verhindern. Die angestrebte GPU-Parallelisierung des KI-Modells soll genau das verhindern und Ressourcenbedarf und Prognosegüte bei gegebener Modellgröße optimieren.

Das Ergebnis

Es wurden **drei Strategien** der GPU-Parallelisierung auf ihre Ressourceneffizienz untersucht: **Tensor Parallelism (TP)**, **Pipeline Parallelism (PP)**, und ihre hybride Kombination, genannt **Hybrid TP & PP**. PP und TP unterscheiden sich in Bezug auf konzeptionelle Umsetzbarkeit, Anzahl der nutzbaren GPUs, Speicherauslastung und Lastverteilung über verschiedene GPUs hinweg. Während PP konzeptionell einfach umsetzbar ist, ist die Anzahl der nutzbaren GPUs limitiert auf die Anzahl der sequentiellen Blöcke mit individueller Rolle im Algorithmus, wie z.B. Encoder, Graph Pooling, Attention Stack und Graph Retrieval. TP greift direkt und kleinteilig in den aufgespannten Tensor ein, lässt sich daher flexibler auf viele GPUs verteilen, ist jedoch aufwändig in der Umsetzung und im Kommunikationsbedarf bzw. der Synchronisierung zwischen GPUs. Die Untersuchung von PP und TP bei Nutzung unterschiedlicher Anzahlen von GPUs ergibt,

Künstliche Intelligenz

High-Performance-Computing

dass die vom KI-Modell generell erzielte Rechenzeiterparnis beibehalten wird, und zusätzlich die vollen 3D-CFD-Datensätze verwendet werden können, wobei TP mehr Potential für zusätzliche Optimierung von Prognosegüte und GPU-Nutzung sowie Rechenzeiterparnis aufweist.

Wie geht es weiter?

Die Ergebnisse bilden die Grundlage für weitere Optimierungen. Zukünftig muss die optimale Anzahl von GPUs für den jeweiligen Anwendungsfall anhand von Performance-Metriken und dem optimalen Verhältnis von Prognosegüte und Netzfeinheit des KI-Modells vor Beginn einer Simulation abgeschätzt werden. Darüber hinaus muss der gesamte Simulationsprozess bestehend aus KI-basierter Prognose mittels GPUs und CFD-basierter Simulation mittels CPUs abgestimmt und vollständig automatisiert werden. Optimierungen des Hybrid TP & PP versprechen zusätzlichen Speedup des Gesamtprozesses.



Ein Projekt der

