ESCADE

Energy-Efficient Large-Scale Artificial Intelligence for Sustainable Data Centers

Programm / Ausschreibung	Digitale Technologien, Digitale Technologien, Digitale Schlüsseltechnologien: Ausschreibung 2022	**Status**	laufend
Projektstart	01.05.2023	**Projektende**	30.04.2026
Zeitraum	2023 - 2026	**Projektlaufzeit**	36 Monate
Keywords	Messwerkzeug, Edge-Cloud Kontinuum, Service Deployment, Nachhaltigkeit, Stahlproduktion		

Projektbeschreibung

Im Projekt ESCADE soll die Energiebilanz von Rechenzentren und speziell KI-Anwendungen durch weltweit führende Hardware- und Software-Technologien wesentlich reduziert werden. Unter anderem wird hierfür eine verteilte KI-Architektur entwickelt und anhand des Anwendungsfalls „Stahlproduktion“ erprobt. Hier wird auch der Fokus des österreichischen Beitrags liegen.

In verteilten cyber-physischen Systemen erfolgt die Datenerfassung bisher oft an den äußeren Rändern des Systems (Edge) und die Verarbeitung in zentralen Rechenzentren mit hoher und effizienter Verarbeitungskapazität (Cloud). An Stelle einer strikten Unterscheidung zwischen Edge und Cloud zeichnet sich jedoch zunehmend ab, dass eine Anwendungs- und Ressourcenabhängige Datenverarbeitung entlang der gesamten Verarbeitungs- und Transportkette sinnvoll sein kann (Edge-Cloud Kontinuum).

Üblicherweise werden Performanz-Kriterien (Rechenperformanz, Kommunikationslatenz) für die Platzierung einer Anwendung im Edge-Cloud Kontinuum angewendet, vernachlässigt wird bisher der Aspekt der Nachhaltigkeit. Es fehlen geeignete Modelle und Lösungsansätze, die in einem potentiell stetigen Veränderungen unterworfenen „Rechen-Kontinuum“ eine intelligente und möglichst automatisierte Platzierung von Diensten und Anwendungen unter Einhaltung notwendiger Performanz-Kriterien bei Minimierung von Umwelt- und Finanzkosten ermöglicht. Eine Lösung dieses multi-dimensionalen Optimierungsproblems wird für die ökologisch-wirtschaftliche Bereitstellung und Nutzung zukünftiger hochverteilter Anwendungen von immenser Bedeutung sein.

In diesem Zusammenhang fehlen auch Messwerkzeuge und Metriken, um die Nachhaltigkeit von verteilten cyber-physischen Systemen über Subsystemgrenzen hinweg zu bewerten. Derartige Messungen sind jedoch Voraussetzung und essentieller Input für etwaige Modellierungs- und Optimierungsansätze.

In diesem Projekt verfolgt die SRFG daher folgende drei primäre Ziele:

1. Die Entwicklung einer validierten Lösung zur Analyse, Bewertung und Empfehlung einer optimierten Platzierung von Diensten und Anwendungen im Edge-Cloud Kontinuum.

2. Die Entwicklung eines Werkzeugs zur automatisierten Messung, Analyse und Visualisierung von Nachhaltigkeitsmetriken.

3. Die Validierung der Ergebnisse aus 1 und 2 im Anwendungsfall „Stahlproduktion“ des Gesamtkonsortiums in einer hochverteilten KI-Architektur.

Das langfristig angestrebte, hochinnovative Ergebnis, für das in diesem Projekt erste Grundlagen und Lösungsansätze geschaffen werden, wird eine vollautomatisierte Lösung sein, die unter Berücksichtigung aller relevanten Kriterien, insbesondere auch Nachhaltigkeit, eine global optimale Platzierung und hochdynamische Migration von Anwendungen innerhalb des Edge-Cloud Kontinuums erlaubt.

Ein weiteres Projektergebnis wird ein Mess-, Analyse-, und Visualisierungswerkzeug sein, mit dem komplexe, geographisch verteilte Systeme über Sub-Systemgrenzen hinweg hinsichtlich ihrer Nachhaltigkeit (zunächst mit Fokus Energieverbrauch) vermessen werden können, um eine Bewertung und Optimierung der Nachhaltigkeit des Gesamtsystems durchführen zu können.

Im Anwendungsfall „Stahlproduktion“ werden diese beiden und weitere Projektergebnisse schließlich zusammengeführt, getestet und validiert, um den Anteil von recyceltem Stahl, der 3,5-mal weniger Energie in der Herstellung benötigt, um bis zu 20% zu steigern.

Nach ersten Schätzungen des Gesamtkonsortiums können basierend auf den Projektergebnissen bis zu 37,5 Mrd. kWh bzw. 5,75 kWh Mrd. pro Jahr an Stromverbrauch in Deutschland bzw. Österreich eingespart und so ein Beitrag zu den umweltpolitischen Zielen und energiepolitischer Unabhängigkeit beider Länder und Europas geleistet werden.

Abstract

In the ESCADE project, the energy consumption of data centers and especially AI applications will be significantly reduced using world-leading hardware and software technologies. Among other things, a distributed AI architecture will be developed and applied to the use case "steel production". The focus of the Austrian contributions to the overall project will also be in this area.

In distributed cyber-physical systems, data acquisition typically takes place at the outer edges of the system (edge) and processing in central data centers with high and efficient processing capacity (cloud). However, instead of a strict distinction between edge and cloud, it is increasingly emerging that application- and resource-dependent data processing can be useful along the entire processing and transport chain (edge-cloud continuum).

Usually, performance criteria (computing performance, communication latency) are used for the placement of an application in the edge-cloud continuum, but the aspect of sustainability has been neglected so far. There is a lack of suitable models and solution approaches that enable intelligent and, if possible, automated placement of services and applications in a "computing continuum" that is potentially subject to constant change, while complying with the necessary performance

criteria and at the same time minimizing environmental and financial costs. A solution to this multi-dimensional optimization problem will be of immense importance for the ecological and economical provision and use of future highly distributed applications.

In this context, there is also a lack of measurement tools and metrics to assess the sustainability of distributed cyber-physical systems across subsystem boundaries. However, such measurements are a prerequisite and essential input for any modeling and optimization approaches.

In this project, SRFG therefore has the following three primary objectives:

1. To develop an approach for analyzing, evaluating, and recommending optimized solutions regarding the placement of services and applications in the edge-cloud continuum.

2. To develop a measurement tool for automated measurement, analysis and visualization of sustainability metrics.

3. The validation of the results from 1 and 2 in the "steel production" use case of the overall consortium in a highly distributed AI architecture.

The long-term, highly innovative result, for which initial foundations and solution approaches will be created in this project, will be a fully automated solution that allows globally optimal placement and highly dynamic migration of applications within the edge-cloud continuum, taking into account all relevant criteria including sustainability in particular.

A further project result will be a measurement, analysis and visualization tool with which complex, geographically distributed systems can be measured across sub-system boundaries with regard to their sustainability (initially with a focus on energy consumption) in order to be able to evaluate and optimize the sustainability of the overall system.

Finally, in the "steel production" use case, these two and other project results will be combined, tested and validated to increase the share of recycled steel, which requires 3.5 times less energy in production, by up to 20%.

According to initial estimates by the overall consortium, up to 37.5 billion kWh or 5.75 kWh per year of electricity consumption can be saved based on the project results in Germany and Austria respectively, thus contributing to the environmental policy goals and energy independence of both countries and Europe.

Projektpartner

- Salzburg Research Forschungsgesellschaft m.b.H.