SwISD

Swarm Inference on Small Devices

Programm / Ausschreibung	Expedition Zukunft, Expedition Zukunft 2023, Expedition Zukunft Start 2023	**Status**	laufend
Projektstart	01.09.2025	**Projektende**	31.08.2026
Zeitraum	2025 - 2026	**Projektlaufzeit**	12 Monate
Keywords	AI, Swarm, Inference, Sustainability, WebGPU,		

Projektbeschreibung

Swarm Inference on Small Devices (SwISD) schlägt die Entwicklung eines dezentralen, webbasierten Peer-to-Peer (P2P)-Netzwerks vor, das die dynamische Zuweisung und Ausführung rechenintensiver Inferenzaufgaben, die in der Regel leistungsstarke Grafikprozessoren (GPUs) erfordern, auf einem heterogenen Ensemble ressourcenbeschränkter Geräte, einschließlich mobiler Geräte und älterer Laptops, ermöglicht. Das Netzwerk bildet einen flexiblen, bedarfsgesteuerten Computerschwarm, in dem ein Kollektiv von Geräten dynamisch zusammengestellt und konfiguriert werden kann, um eine skalierbare und anpassungsfähige Ressource für künstliche Intelligenz (KI) bereitzustellen. Diese auf Schwärmen basierende Architektur erinnert an kompartimentierte R1-Reasoning-Frameworks für große Sprachmodelle (LLMs), bei denen komplexe Aufgaben in kleinere, modulare Komponenten zerlegt werden, die parallel über ein verteiltes Netzwerk ausgeführt werden können. Durch die Nutzung der aggregierten Verarbeitungskapazitäten des Schwarms kann das Netzwerk ein flexibles und effizientes Mittel zur Ausführung von KI-Arbeitslasten bieten, das die Schaffung eines skalierbaren, fehlertoleranten und selbstorganisierenden Systems ermöglicht, das sich an veränderte Rechenanforderungen anpassen kann. Die dezentrale und dynamische Natur des Schwarms ermöglicht es dem Netzwerk, die Ressourcennutzung zu optimieren, die Latenzzeit zu minimieren und den Durchsatz zu maximieren, was es zu einer attraktiven Lösung für eine breite Palette von KI-Anwendungen macht, von Natural Language Processing bis zu Computer Vision und darüber hinaus.

Abstract

Swarm Inference on Small Devices (SwISD) proposes the development of a decentralized, web-based peer-to-peer (P2P) network that enables the dynamic allocation and execution of computationally intensive inference tasks, typically necessitating high-performance graphics processing units (GPUs), on a heterogeneous ensemble of resource-constrained devices, including mobile devices and legacy laptops. The network forms a flexible, on-demand computing swarm, wherein a collective of devices can be dynamically assembled and configured to provide a scalable and adaptive resource for artificial intelligence (AI) workloads. This swarm-based architecture is reminiscent of compartmentalized R1 reasoning frameworks for large language models (LLMs), where complex tasks are decomposed into smaller, modular components that can be executed in parallel across a distributed network. By harnessing the aggregate processing capabilities of the swarm, the network can provide a flexible and efficient means of executing AI workloads, allowing for the creation of a scalable, fault-

tolerant, and self-organizing system that can adapt to changing computational demands. The decentralized and dynamic nature of the swarm enables the network to optimize resource utilization, minimize latency, and maximize throughput, making it an attractive solution for a wide range of AI applications, from natural language processing to computer vision and beyond.

Projektpartner

- Homahuki GmbH