Kelis

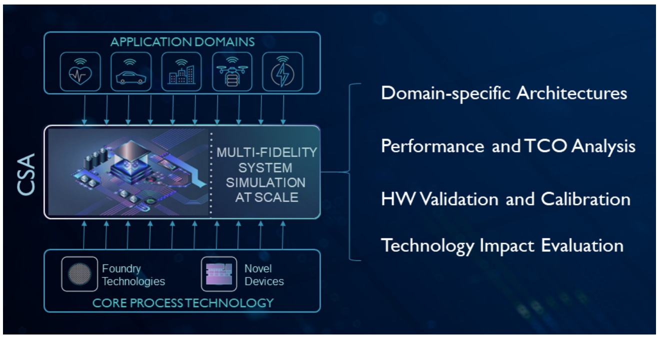
Kelis: performance, power, area and cost modeling of AI datacenters built for large language model training and inference

Large language models (LLMs) have proven to be of high value and show an even larger promise for future applications in numerous areas of science and technology. Many innovations stem from scaling the complexity of those models, requiring an **ever-increasing computational need** from the systems that execute them. Uniquely positioned in the R&D landscape across the entire system and technology stack, imec aims to solve scaling issues by creating new approaches to computing, connectivity, and architecture for future high-performance systems.

Compute System Architecture (CSA) is a center of excellence in imec enabling **true hardware-software-technology codesign to architect HPC and AI systems of the future**. The goal is to pathfind energy-conscious, high-performance computing solutions at scale for the chiplet era. The team analyzes emerging usage models and architects compute systems capable of post-exascale performance. Imec's expertise in system-level modelling, performance analysis and hardware validation enables system exploration encompassing the entire spectrum of workload complexity, modelling granularity and technology maturity.

Imec's Kelis tool provides fast and accurate performance, power, area and cost evaluation and design space exploration for AI datacenters running large language models, helping to quickly evaluate and optimize design choices. **Validated within 12% worst case error** for large scale LLM training and inference executions on Nvidia A100 and H100 systems, Kelis returns results within seconds, allowing for interactive exploration. The framework leverages imec's expertise in analytical performance modeling for high-performance computing and artificial intelligence.

Imec, the world-leading independent R&D centre on semiconductor technology, headquartered in Leuven, Belgium, hosts over 6,000 experts from 100 countries.


APPLICATION DOMAINS
CSA
MULTI-FIDELITY SYSTEM SIMULATION AT SCALE
Foundry Technologies
Novel Devices
CORE PROCESS TECHNOLOGY
Domain-specific Architectures
Performance and TCO Analysis
HW Validation and Calibration
Technology Impact Evaluation

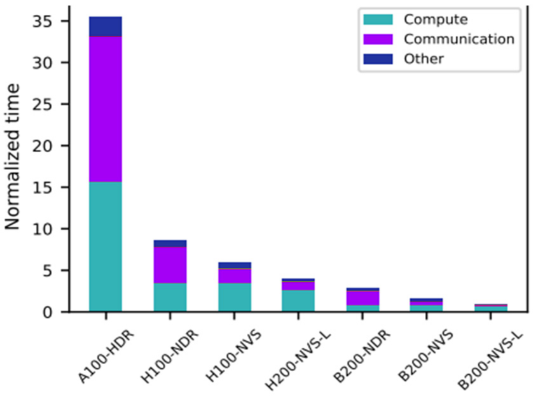
CSA – imec centre of excellence for HW-SW codesign of future compute systems, enables novel system architectures using scalable multi-fidelity simulation framework.

Use Kelis to optimize AI datacenter performance/TCO

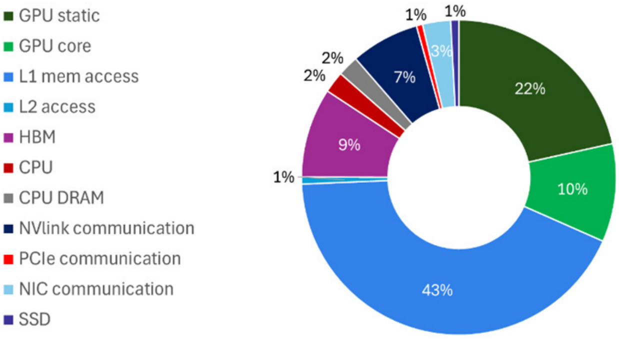
The Kelis tool offers modeling of the compute, communication, and memory subsystem of accelerators, as well as the larger systems they are part of, all the way to complete datacenter scale. Kelis consists of an LLM task-graph analyzer, a parallelism mapper, a hierarchical roofline model, a topology-aware collective communication library along with many knobs to play with, that give the user an **end-to-end framework for all the explorations regarding large-scale AI datacenters and relevant workloads**. The Kelis framework is capable of evaluating the impact of new compute, memory, or communication technology for AI workloads.

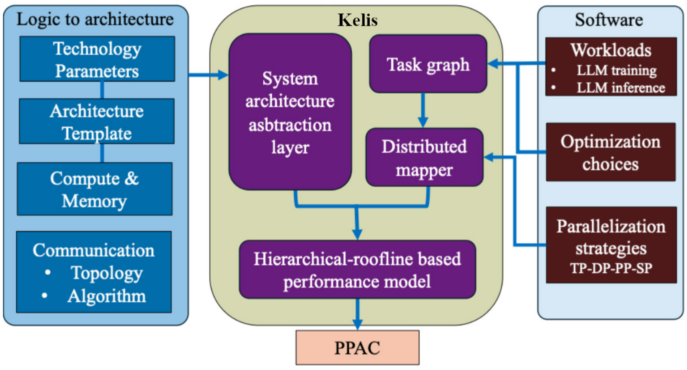
Contact us for more information and licensing.

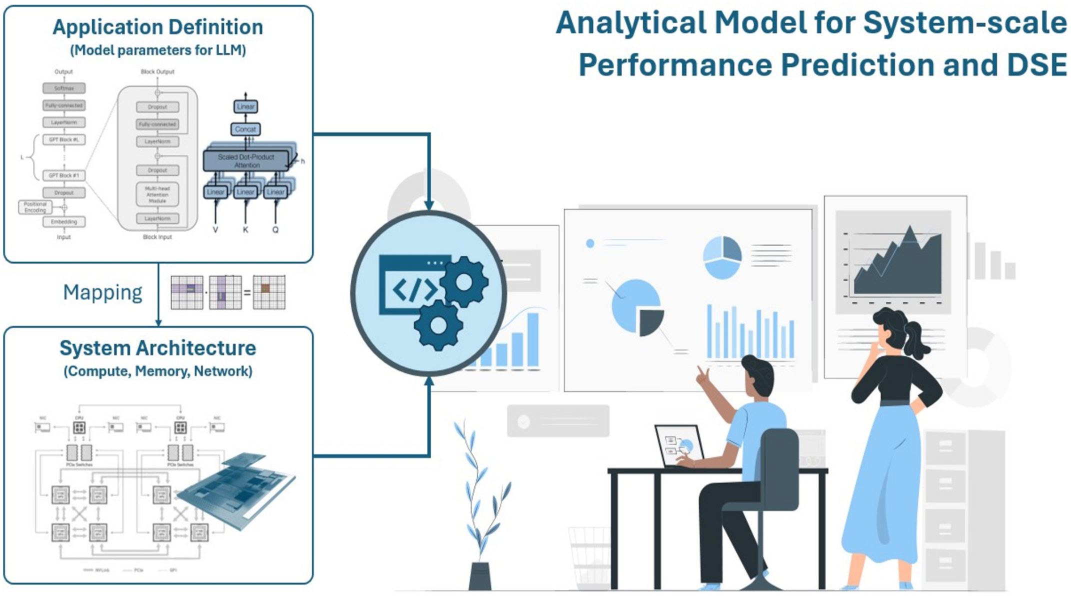
Normalized training performance scaling across GPU generations and networking systems for GPT3-175B



Energy breakdown of a single node during training of GPT3-175B on a system representing a 16k A100 cluster






Application Definition
(Model parameters for LLM)
Mapping
System Architecture
(Compute, Memory, Network)
Analytical Model for System-scale
Performance Prediction and DSE

Kelis offers an easy-to-use, interactive interface exposing key parameters, showing results on a dynamic dashboard.

Reference: J. Kundu, W. Guo, A. BanaGozar, U. De Alwis, S. Sengupta, P. Gupta & A. Mallik, (2024). Performance Modeling and Workload Analysis of Distributed Large Language Model Training and Inference. arXiv preprint arXiv:2407.14645 (IISWC'24).

FIND OUT MORE:

CONTACT US
WWW.CONTACTIMEC.COM

DISCLAIMER - This information is provided 'AS IS', without any representation or warranty.Imec is a registered trademark for the activities of IMEC International (a legal entity set up under Belgian law as a 'stichting van openbaar nut', Register of Legal Entities Leuven, VAT BE 0817.807.097), imec Belgium (Interuniversitair Micro-Electronica Centrum vw supported by the Flemish Government, Register of Legal Entities Leuven, VAT BE 0425.260.668), imec the Netherlands (Stichting IMEC Nederland, part and co-founder of Stichting Holst Centre which is supported by the Dutch Government and known in the Dutch Kamer van Koophandel under the number 17179812), imec Taiwan (IMEC Taiwan Co. registered under the business license number 28112596), imec China (IMEC Microelectronics (Shanghai) Co. Ltd. registered under the business license number 310115400259925), imec India (IMEC India Private Limited registered under corporate number U73100DL2011FTC222076), imec San Francisco office (IMEC Inc. registered under corporate number N16000004843).