

IEEE ASAP 2026 · TUTORIAL

The AIxSIM Project

Open-source instruments for cross-stack simulation of AI models on new hardware.

**Imperial College
London**



THE UNIVERSITY
of EDINBURGH

2 September 2026 14:30 – 16:00

Funded by ARIA · Scaling Compute

Binglei Lou

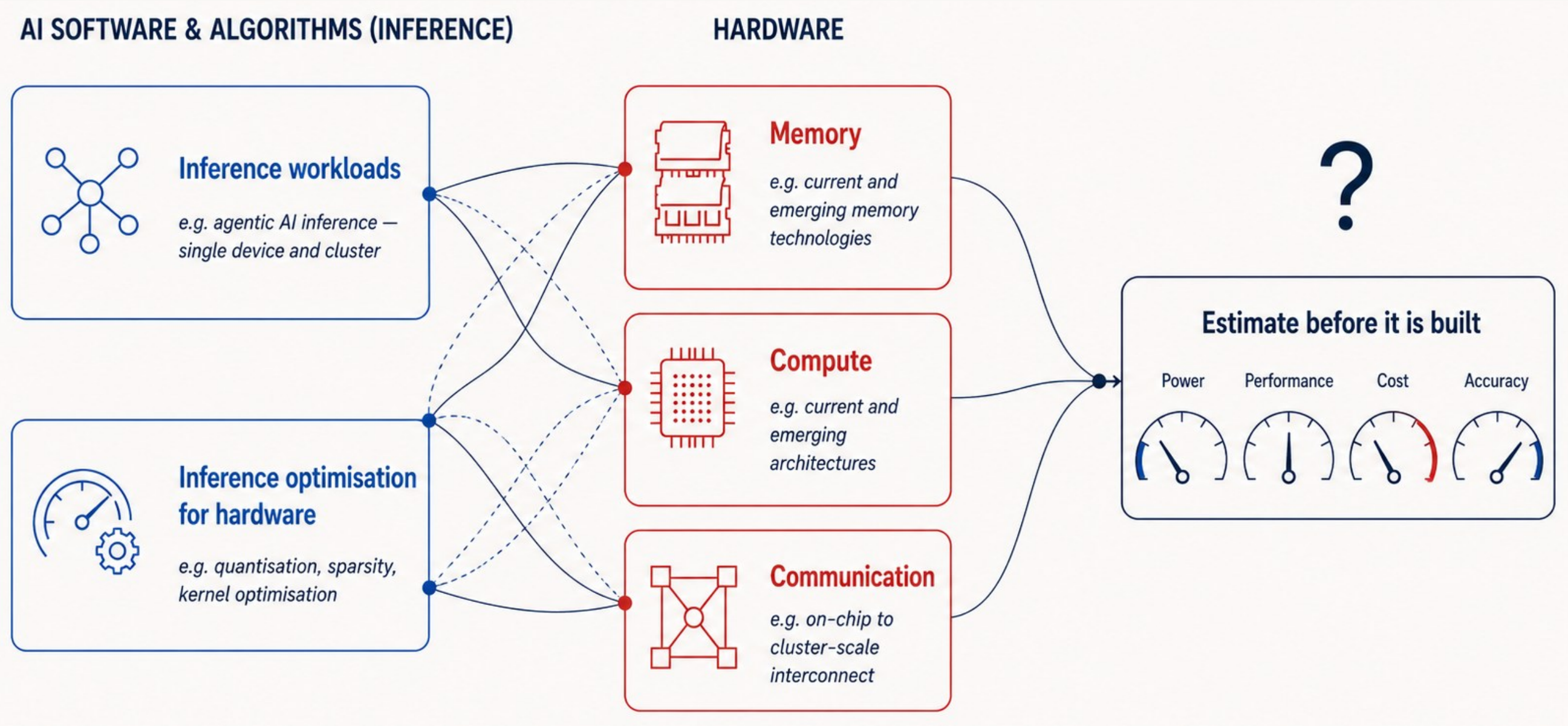
Imperial College London · Tutorial
Chair

aicrosssim.github.io

[2026.asap-conf.org/aixsim-
project-tutorial](https://2026.asap-conf.org/aixsim-project-tutorial)



Emerging algorithms on hardware that does not exist yet



RTL simulation

Most accurate, slowest — hours per layer

Transaction-level

Block behaviour, no cycle detail

Analytical

Seconds to minutes, cluster scale

Funded by ARIA (UK) · Scaling Compute programme

aicrosssim.github.io/research

THIS AFTERNOON

Three talks, from one die to the cluster

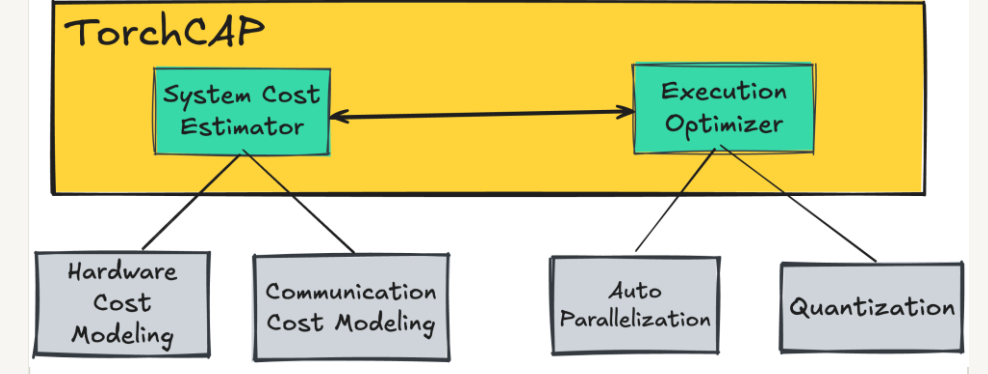
aicrosssim.github.io/research



single die



code for that die



many dies, one cluster

01 SINGLE DIE

PLENA

Accelerator architecture against the memory walls of long-context LLM inference.

George Haoran Wu
University of Cambridge

02 CODE FOR THAT DIE

KernelCraft

Can agents write close-to-metal kernels for AI hardware? Benchmarks, harnesses, self-improvement.

Jiayi Nie
University of Cambridge

03 CLUSTER OF DIES

SysSim

Step time, MFU and per-GPU memory for distributed LLM training — on hardware you do not have.

Luo Mai
University of Edinburgh