Hespas: Multi-fidelity simulation of StableHLO-ML workloads with ASTRA-sim

Abubakr Nada
Imec

imec

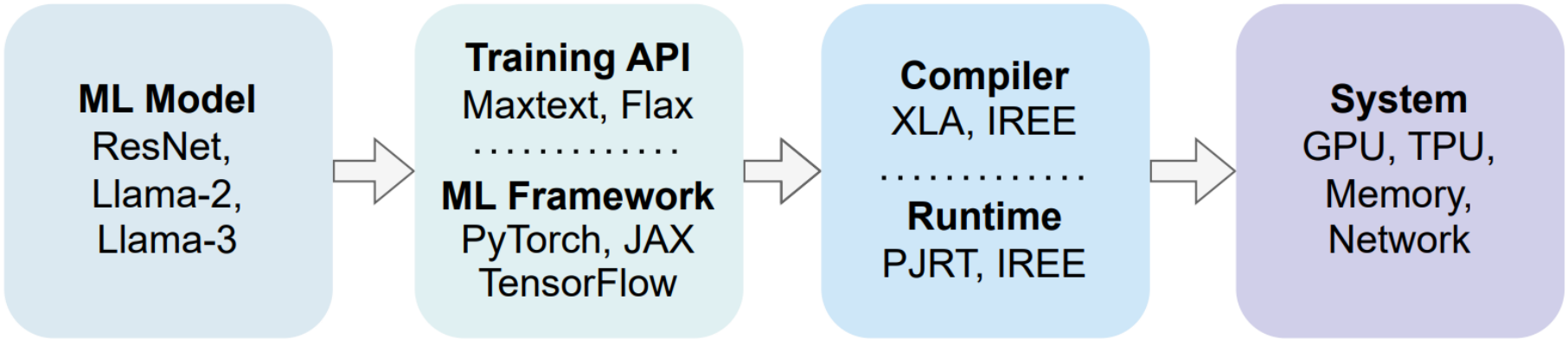
ISCA 2026, Raleigh NC

This work is funded by the Advanced Research + Invention Agency (ARIA).

Challenges in Distributed ML Training

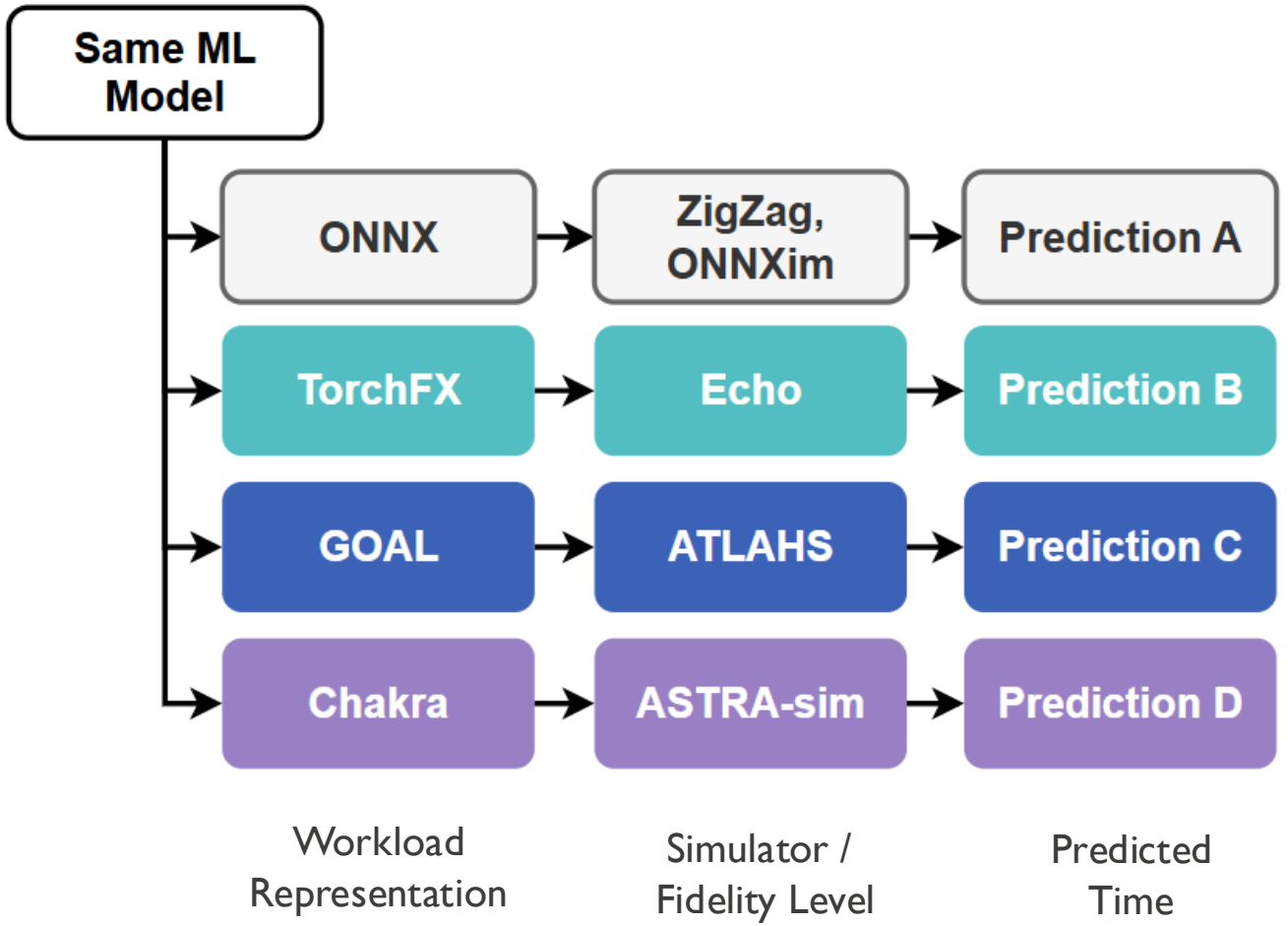
Co-optimization across the ML stack

- ML training performance prediction is a multi-dimensional, cross-stack problem
- Empirical evaluation is too costly
- Simulators are key drivers for innovation



Fragmentation in Distributed ML simulators

- ML performance modeling uses **multiple tools** and **fidelities**
- Each relies on **different workload representations**
- Often the **same model** uses a different workload representations
- Representations are **not portable** or **forward-compatible**

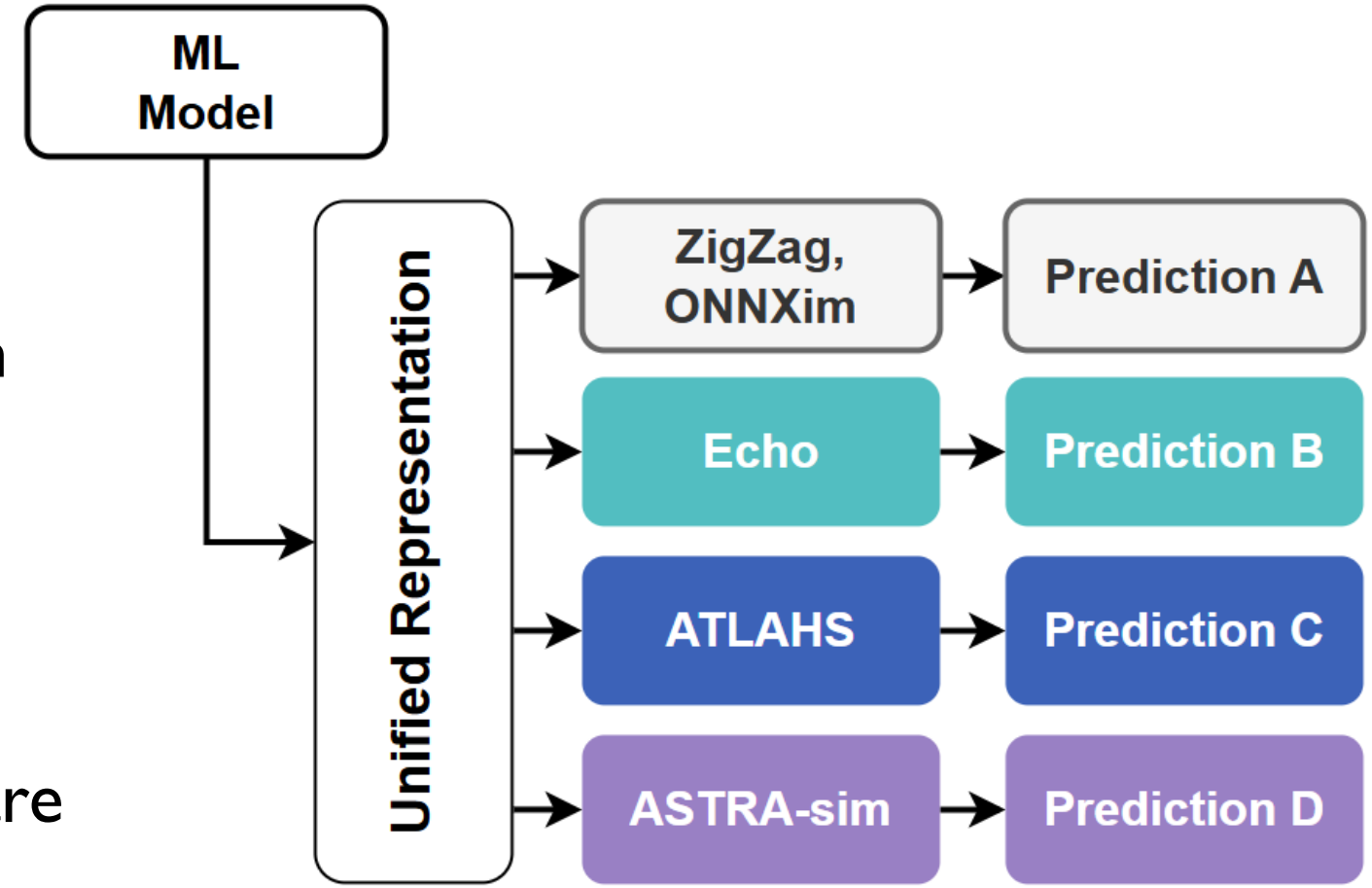


Difficult to **compare, validate, and reuse** workloads.

imec

Requirements for a Unified Workload Representation

- **Model** Agnostic
- **Hardware** Agnostic
- Captures **Communication**
- Supports multiple **fidelities**
- Ahead of time, **Portable**
- **Executable** on real hardware



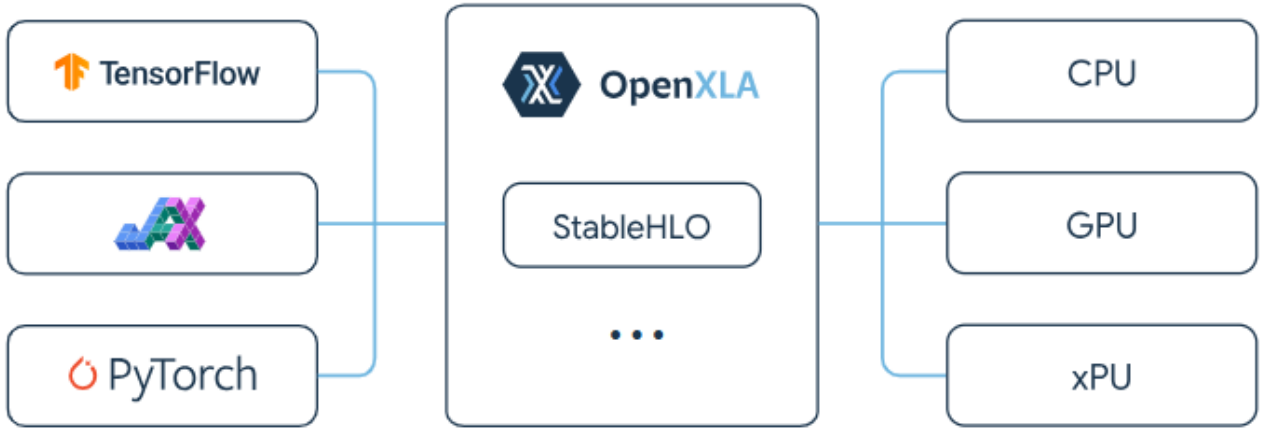
We choose StableHLO as a unified workload representation

https://openxla.org/stablehlo/spec

What is StableHLO?

MLIR

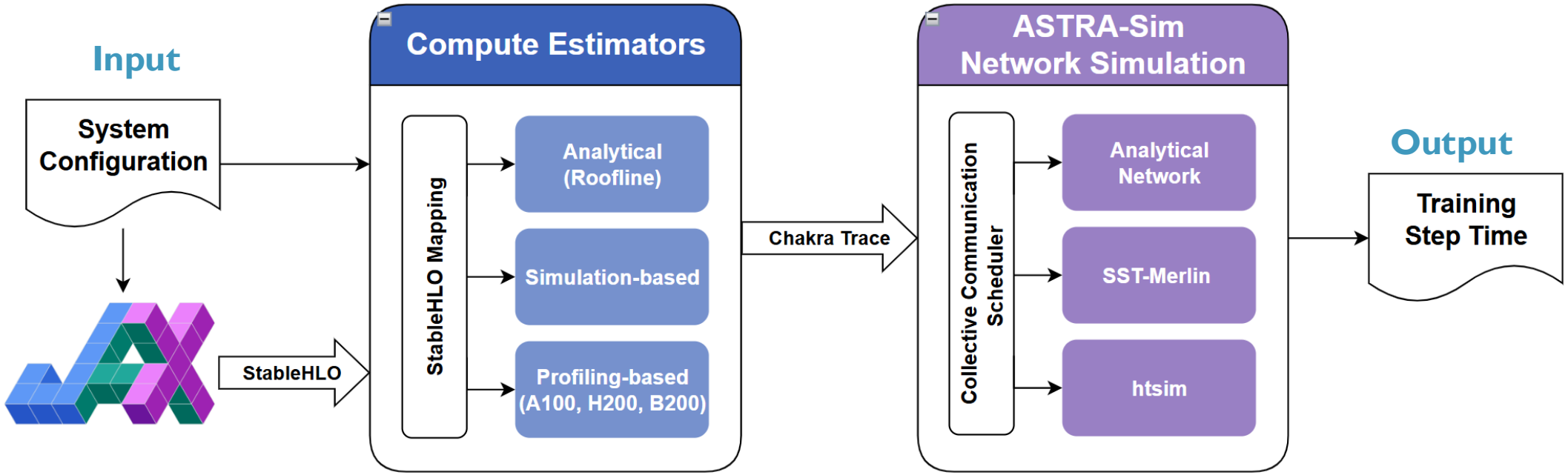
- ML **operator-level** IR (~100 ops.)
  - Compute + collectives (e.g. matmul, all_reduce)
- **Framework** Agnostic
  - **JAX, Pytorch** via **OpenXLA**
- **Ahead-of-time** export
- **Compilable** and **executable**
- **Portable** representation (GPU, TPU, CPU flows)



StableHLO is high-level enough for simulation,
yet concrete enough for execution

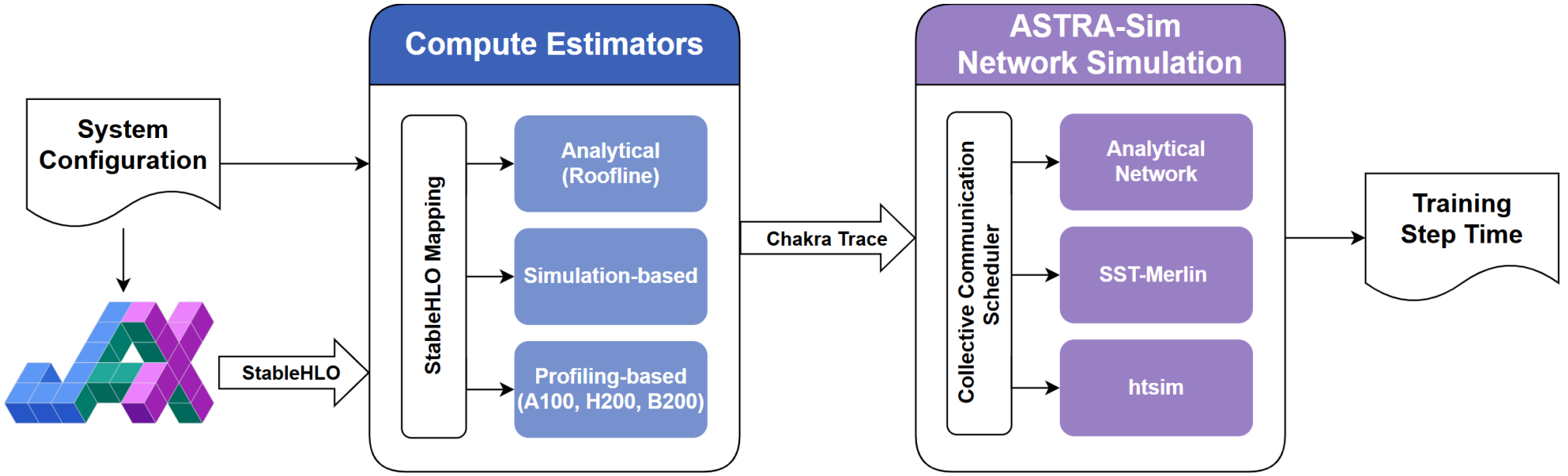
Hespas Simulation Framework

Hespas Simulation Framework Overview

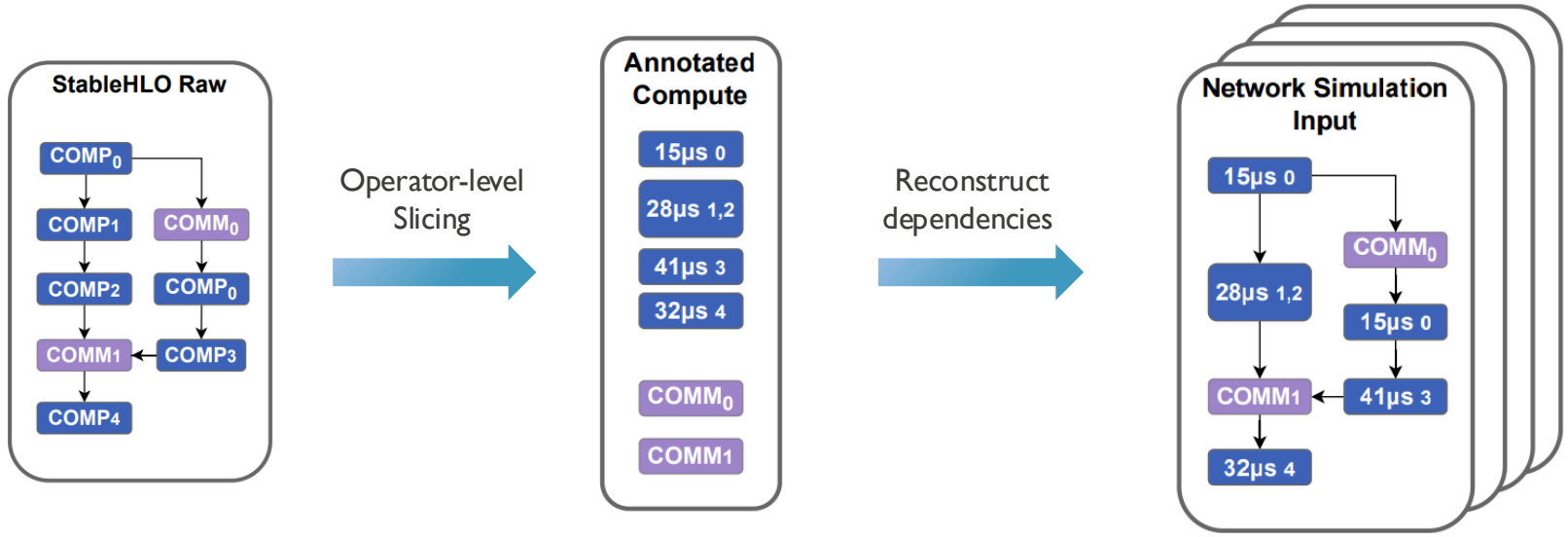


Performance Simulation Pipeline

Simulation Pipeline



Workload transformation Pipeline

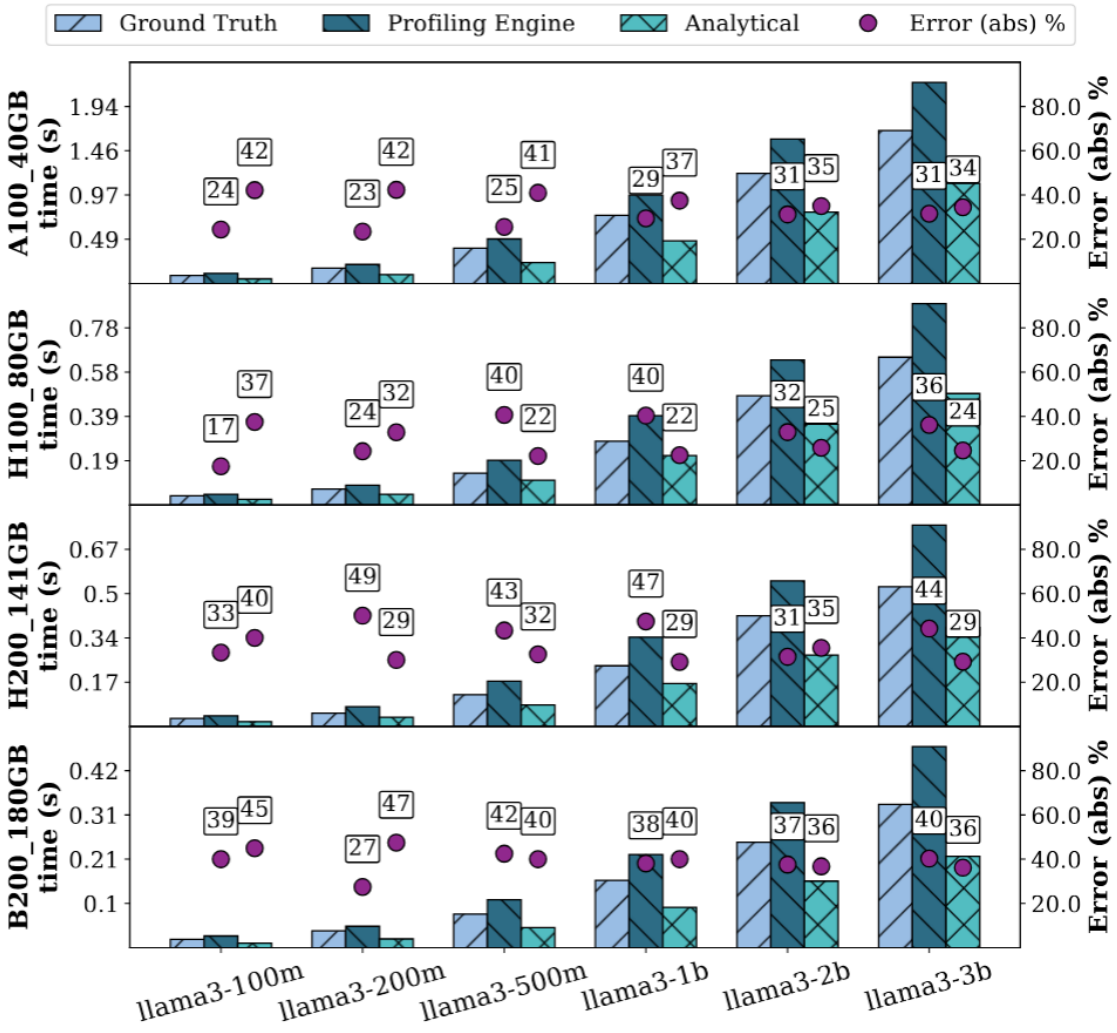


imec

Experimental Setup

- **Export**
  - JAX / **MaxText**
  - **Single program** for all devices (SPMD)
- **Workloads**
  - ResNet (18-200) (4 GPUs DP)
  - Llama2 (7B) (16-128 GPUs DP)
  - Llama3 (100M-3B) (4 GPUs FSDP)
- **Systems**
  - **GPUs**: A100, H100, H200, B200
  - **TPUv3** (8 cores)
- **Optimization**
  - OpenXLA (**hlo-opt** pipeline)
- **Evaluated Models**:
  - **Analytical**-based (GPU, TPU)
  - **Profiling**-based (GPU)
  - **Simulation**-based (TPU)

4xA100 System Simulation Llama-3



Analytical < *Ground Truth* < Profiling

Optimistic
22-42% Error

Pessimistic
17-44% Error

Transition	Profiling speedup Error	Analytical speedup Error
A100 – H100	3%	15%
H100 – H200	7%	-7%
H200 – B200	-3%	14%

Relative speedup error is computed, for example:

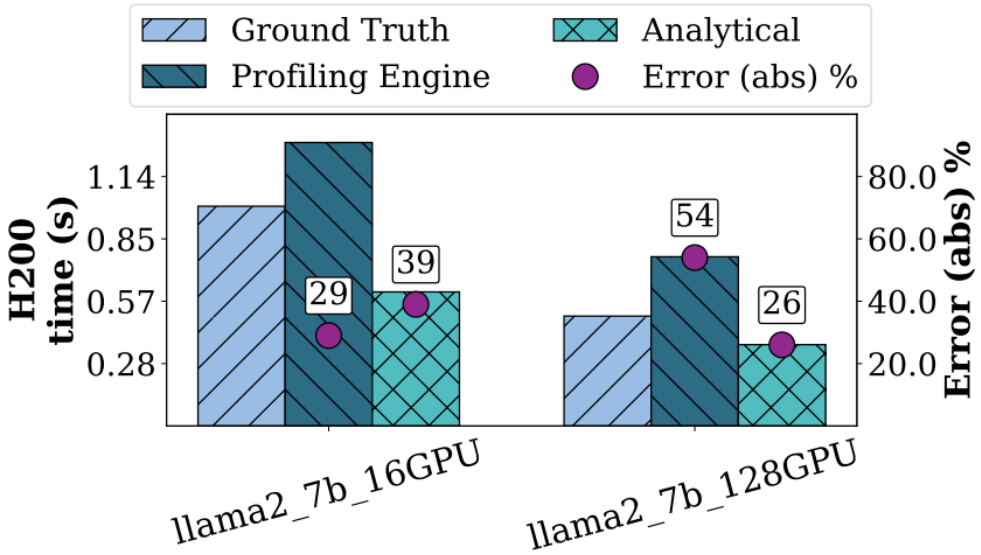
$$\frac{S_{Ref} - S_{Prof}}{S_{Ref}}, \text{where } S = \frac{T_{A100}}{T_{H100}}$$

Analytical underestimates, profiling overestimates, but trends are preserved.

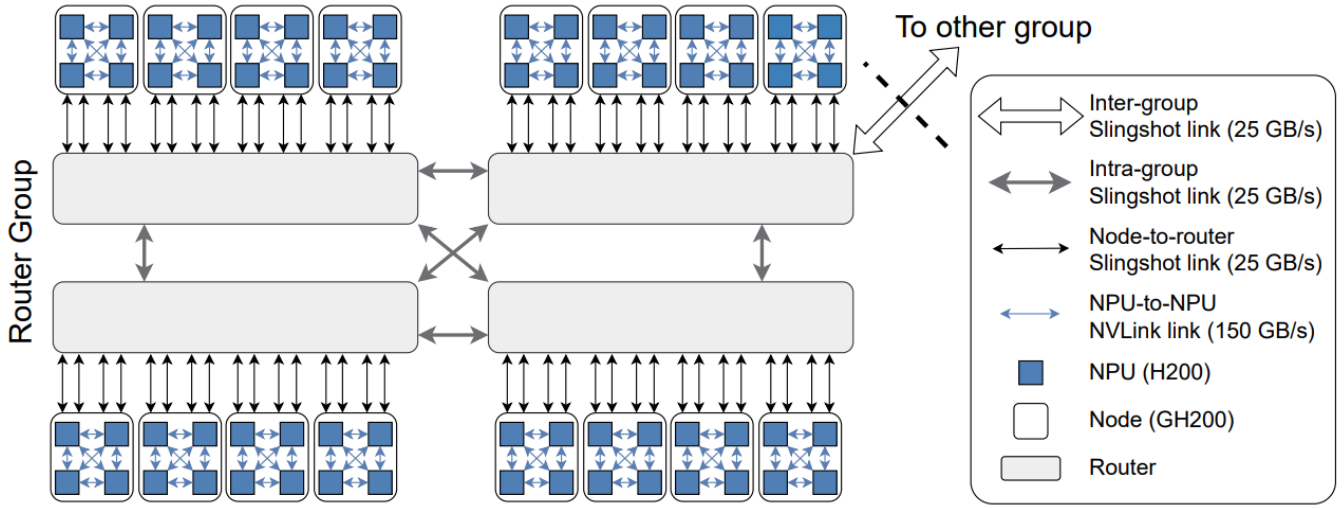
imec

16-128 H200 Scale-out System Simulation

No large-scale cluster needed for simulation



Ground Truth correspond to runs from ATLAHS* paper, run on the Alps supercomputing cluster.

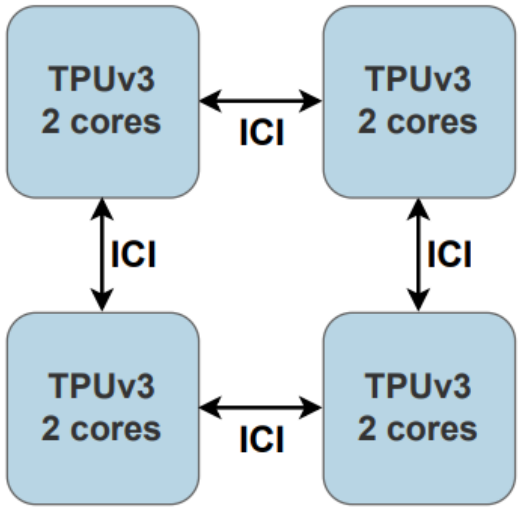


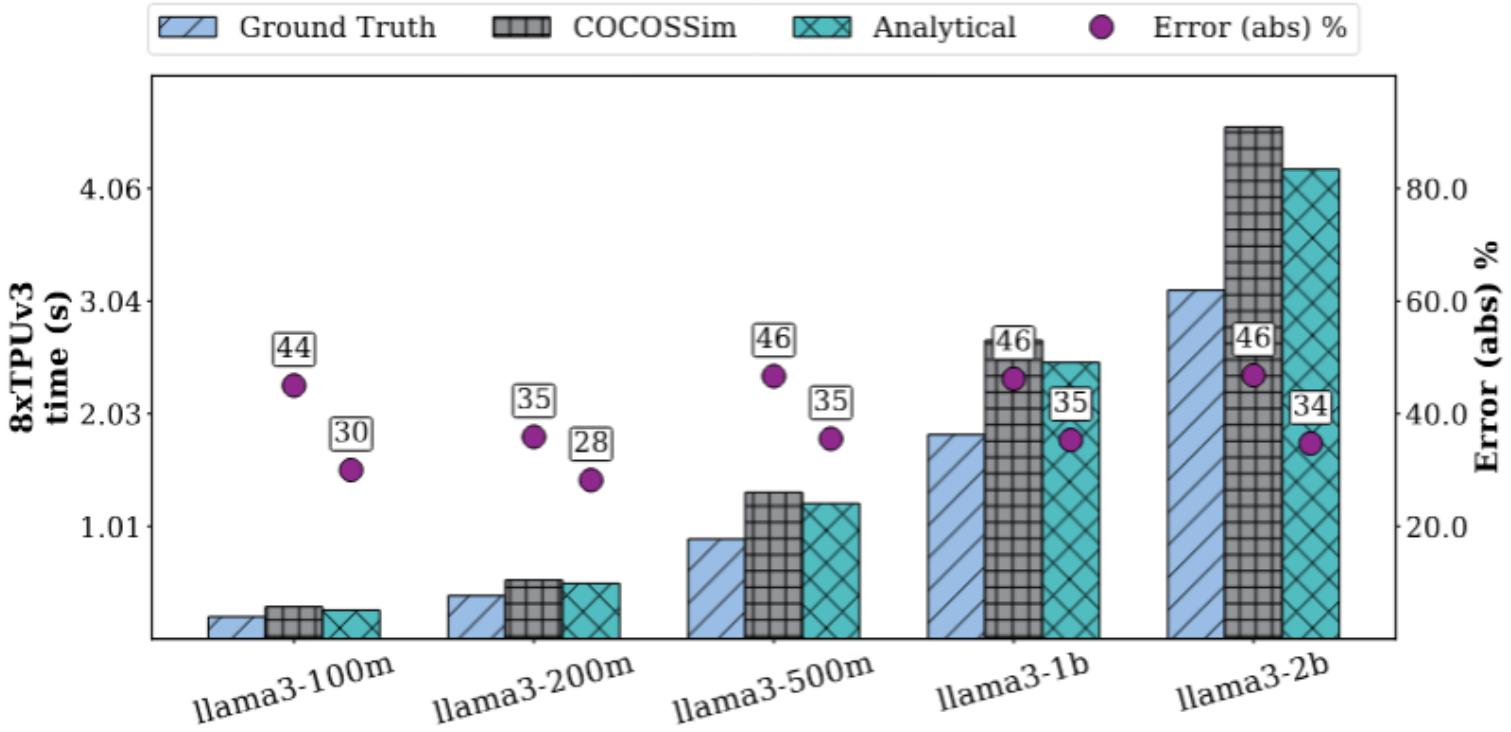
Dragonfly Network Topology
(Modelled via SST-Merlin†)

Performance trends are preserved as system scale.

*https://arxiv.org/abs/2505.08936. † https://sst-simulator.org/sst-docs/docs/elements/merlin/intro

4 x TPUv3 System Simulation with Llama3





Analytical and simulation-based prediction underestimates TPU performance due to lower control over TPU-specific compiler optimizations.

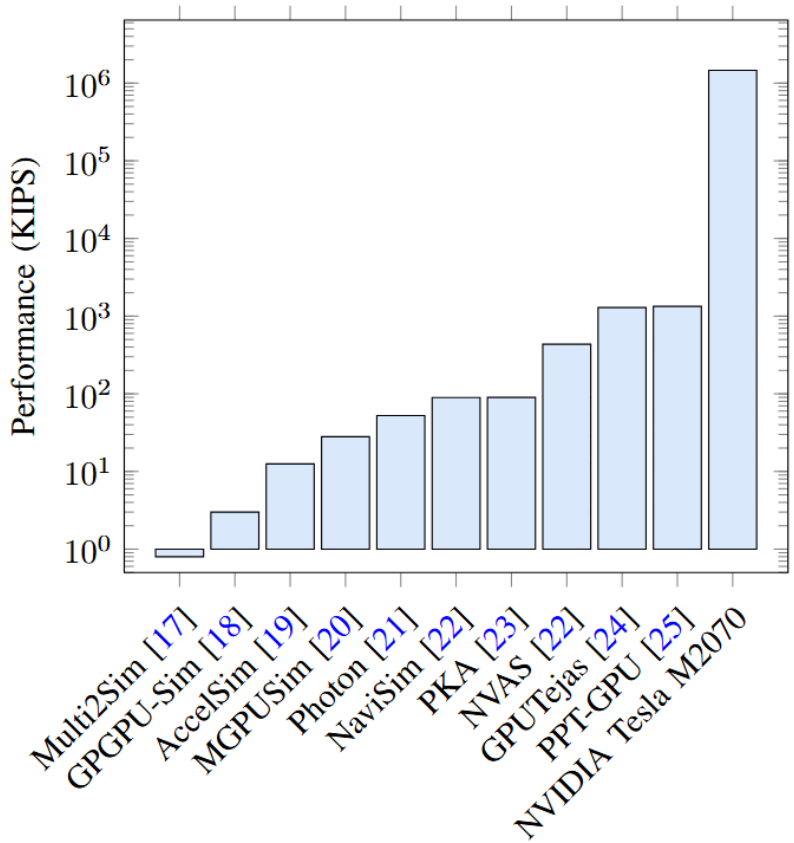
imec

Open Challenges

- Integrating detailed Compute Simulator is challenging
  - Complex integration flows
  - Low simulation rate for ML

- User defined Compute Estimators may not always be able to cleanly map to existing

- Limited control over TPU optimization passes (closed-source OpenXLA for TPUs)



GPU Models' Simulation Rate

https://arxiv.org/abs/2506.09275

Key Takeaway

One Workload Representation Multiple **ML Models**, **Hardware** and **Simulators**

Workloads

- Resnet 18 – 200 (4 GPU)
- Llama2 7B (4-128 GPU)
- Llama3 100M-3B (4-8 GPU)

Systems Configurations

- NVIDIA: A100, B200, H100, H200
- AMD MI300X
- TPUv3

Estimators via Hespas Compute API

- **Analytical**
  - Roofline
- **Detailed**
  - COCOSSim
  - ONNXim
  - ZigZag
  - SCALE-sim
- **Profiling**
  - IREE
  - XLA

Workloads: https://github.com/imec-int/hespas_workloads
Estimators: https://github.com/imec-int/hespas

Thank you

hespas

github.com/imec-int/hespas

ISPASS 2026
Tutorial