Attention: Tutorial is being recorded

AI System Tradeoff & Resource Analysis simulator

Enabling Software-Hardware Co-Design Exploration for Distributed AI Platforms

ASTRA-sim Tutorial
@ISCA 2026
June 27, 2026

https://astra-sim.github.io/tutorials/isca-2026

https://astra-sim.github.io

Welcome

Organizing Team

Tushar Krishna
Associate Professor
Georgia Institute of Technology

Brad Beckmann
Fellow
AMD Research

William Won
Post-Doctoral Researcher
AMD Research

Tuan Ta
Member of Technical Staff
AMD Research

Jinsun Yoo
PhD Candidate
Georgia Institute of Technology

Invited Speakers!

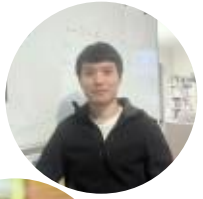


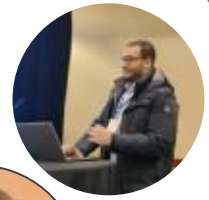










AI is pervasive today!

Chatbots

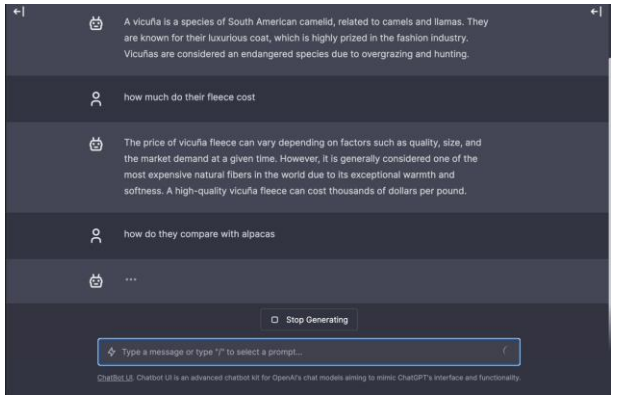
Code Generation

```ruby
def segmentation(items, separator)
  curr = []
  segments = []
  items.each do |item|
end

segmentation([1,2,4,0,2,5,0,3,0], 0).each do |segment|
  puts(segment.join(", "))
end
```

"25% of code at google in last quarter was AI generated"

Text Generation



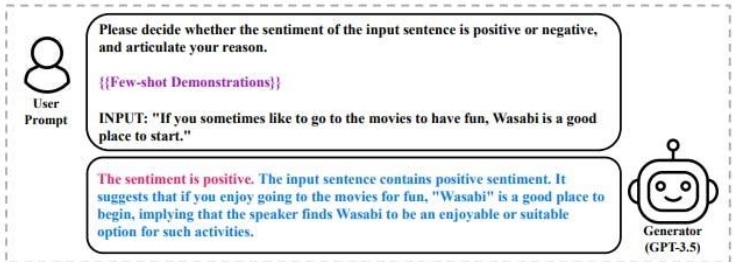
Language Translation

[*Instruction*]: Translate the following sentences from English to Chinese.
[*Input*]: Did you see it go?

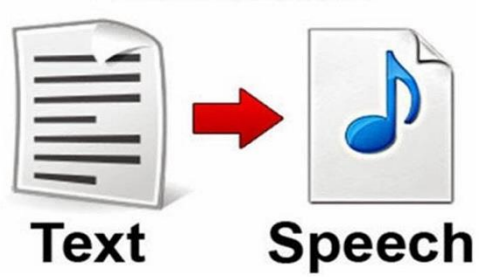
LLM

[*Output*]: 看清楚了吗？

Sentiment Analysis

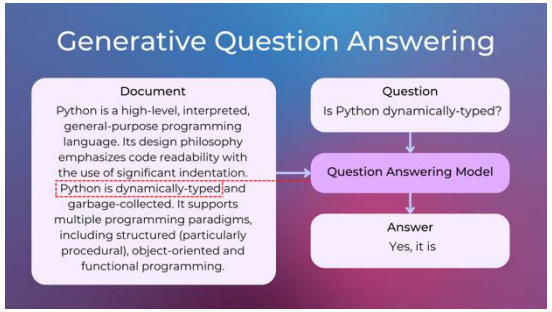


Text to Speech



Recommendations

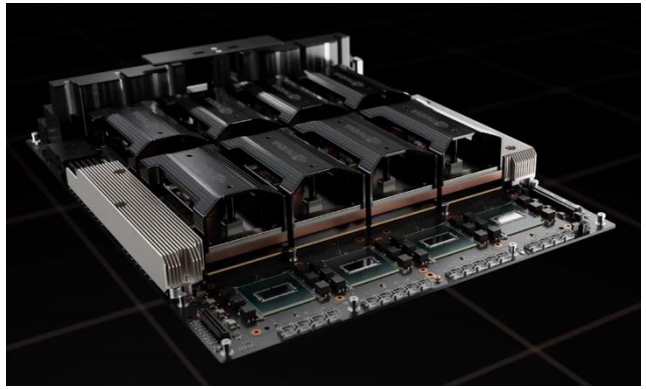
Question Answering



Algorithmic view of AI (Datasets and Models)

Computer Architect's view of AI

"AI Datacenters"
"AI Supercomputers"



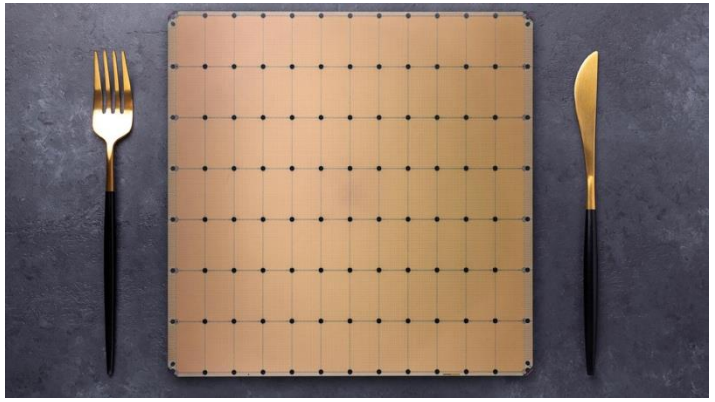
NVIDIA
HGX-H100 SuperPod

Google Cloud
TPUv4

SambaNova
SN40L

AMD
Instinct Platforms

Intel
Gaudi

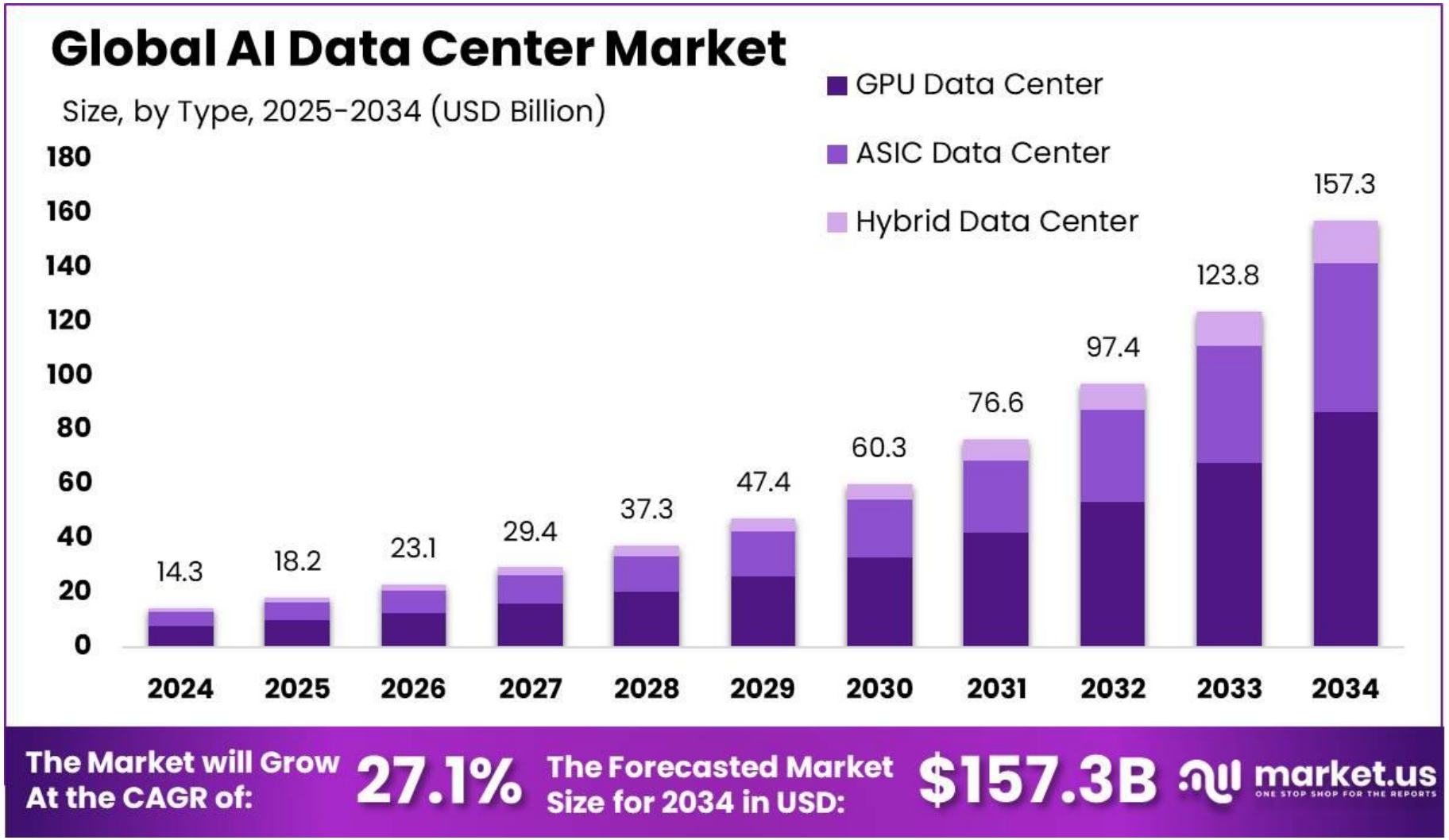


Cerebras
Andromeda

AI is a distributed systems problem!

	GPT-4 (2023)	GPT-5 (2025)
Total Parameters (Weights)	~1.8 Trillion (using MoE)	Unknown
Training Compute	~25000 NVIDIA A100 GPUs over 90-100 days	~170,000 H100/H200 GPUs over 2 years
Training Data	~13 Trillion Tokens	~70 Trillion Tokens
Inference Compute	128 NVIDIA A100 GPUs	Multiple GB200 (72 GPU) nodes
Context Length	32,000 Tokens	400,000 Tokens

The AI datacenter market continues to grow!



The AI datacenter "scale" continues to grow!

Google The Keyword Home Product news Company news Feed

TECHNOLOGY › RESEARCH

Nov 04, 2025

Meet Project Suncatcher, a research moonshot to scale machine learning compute in space.

Artificial intelligence is a foundational technology that could help us tackle humanity's greatest challenges. Now, we're asking where we can go next to unlock its fullest potential. Today we're announcing Project Suncatcher, our new research moonshot to one day scale machine learning in space. Working backward from this potential future, we're exploring how an interconnected network of solar-powered satellites, equipped with our Tensor Processing Unit (TPU) AI chips, could harness the full power of the Sun.

Inspired by other Google moonshots like autonomous vehicles and quantum computing, we've begun work on the foundational work needed to one day make this future possible. We're excited that this is a growing area of exploration, and our initial research, shared today in a preprint paper, describes our approach to satellite constellation design, control, and communication, and also our initial learnings from radiation testing Google TPUs.

Our next step is a learning mission in partnership with Planet to launch two prototype satellites by early 2027 that will test our hardware in orbit, laying the groundwork for a future era of massively-scaled computation in space.

We are having to invent new terminology!

Scale-In → Scale-Up → Scale-Out → Scale-Across → Scale-Above

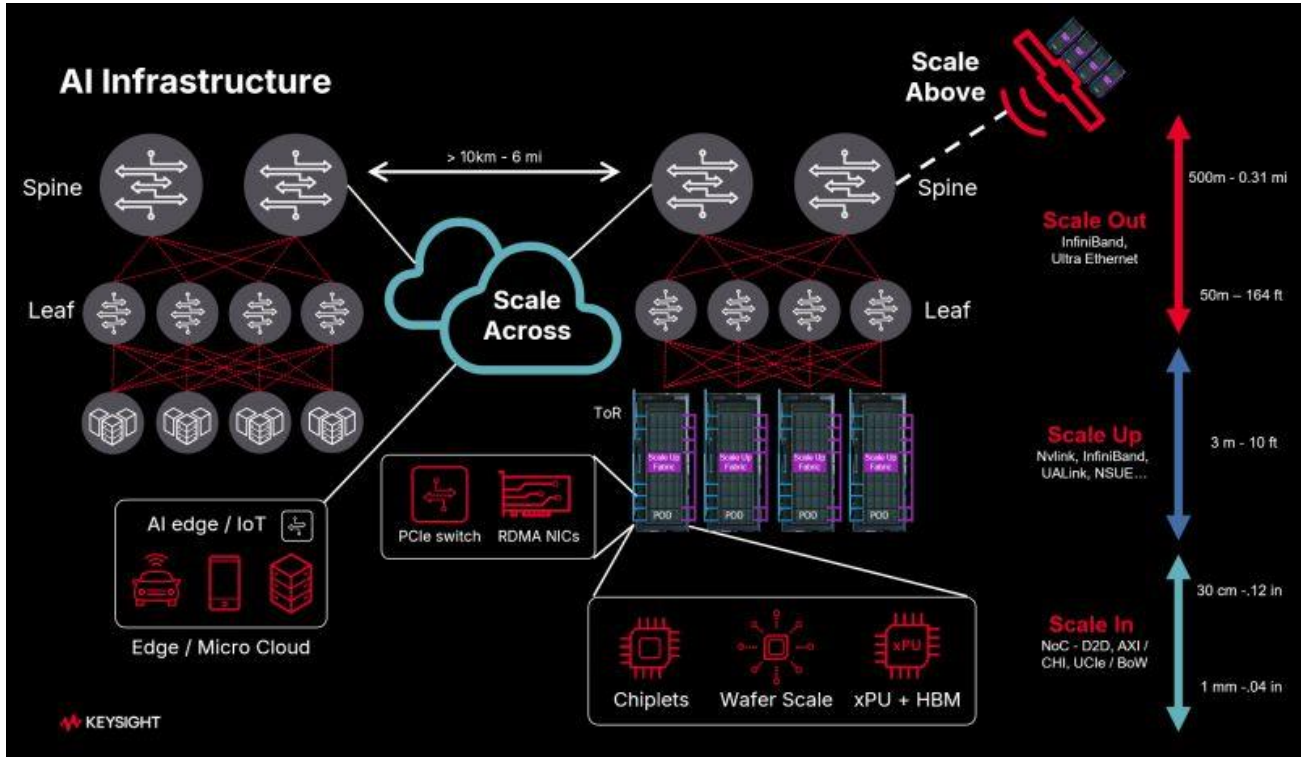
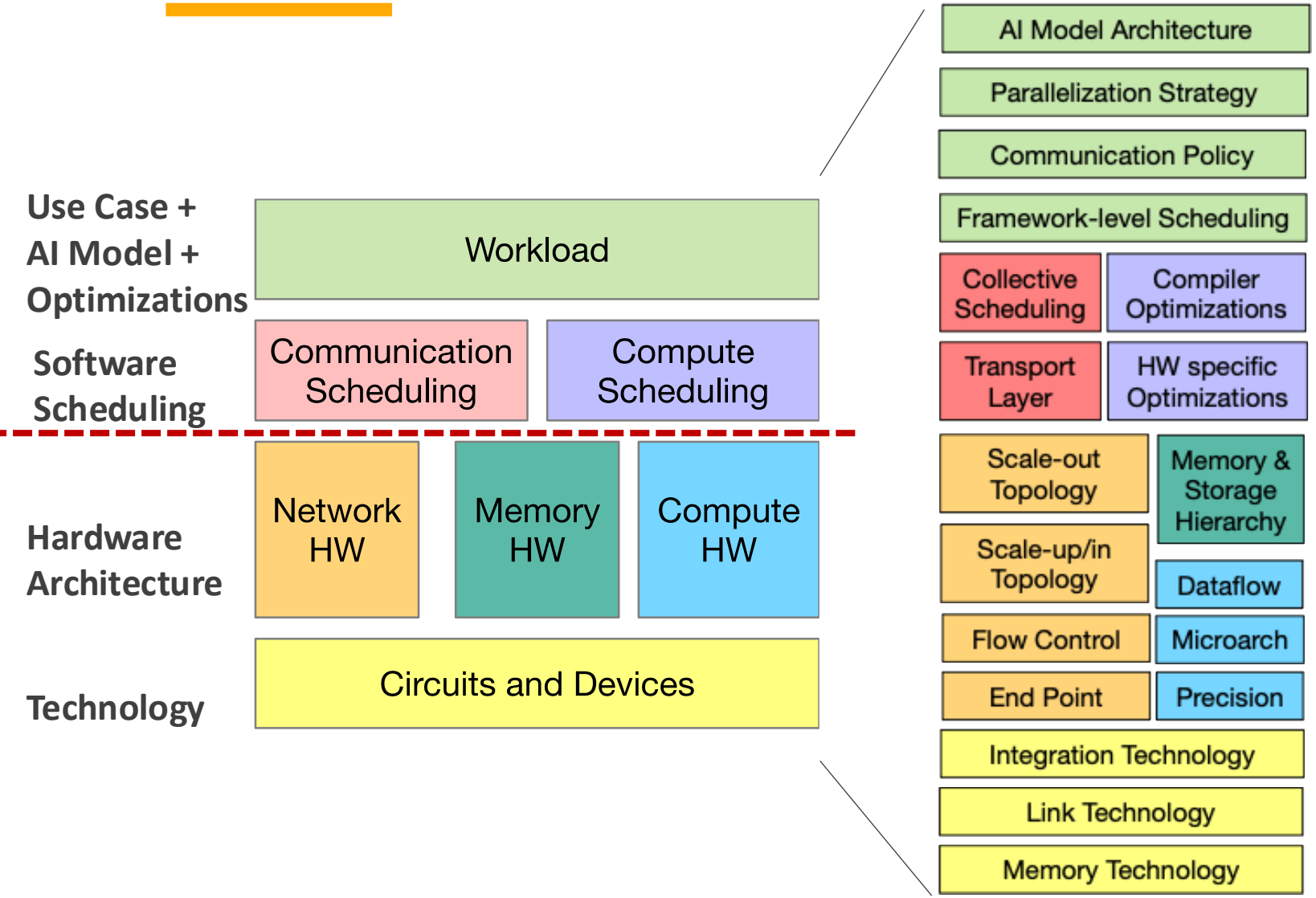


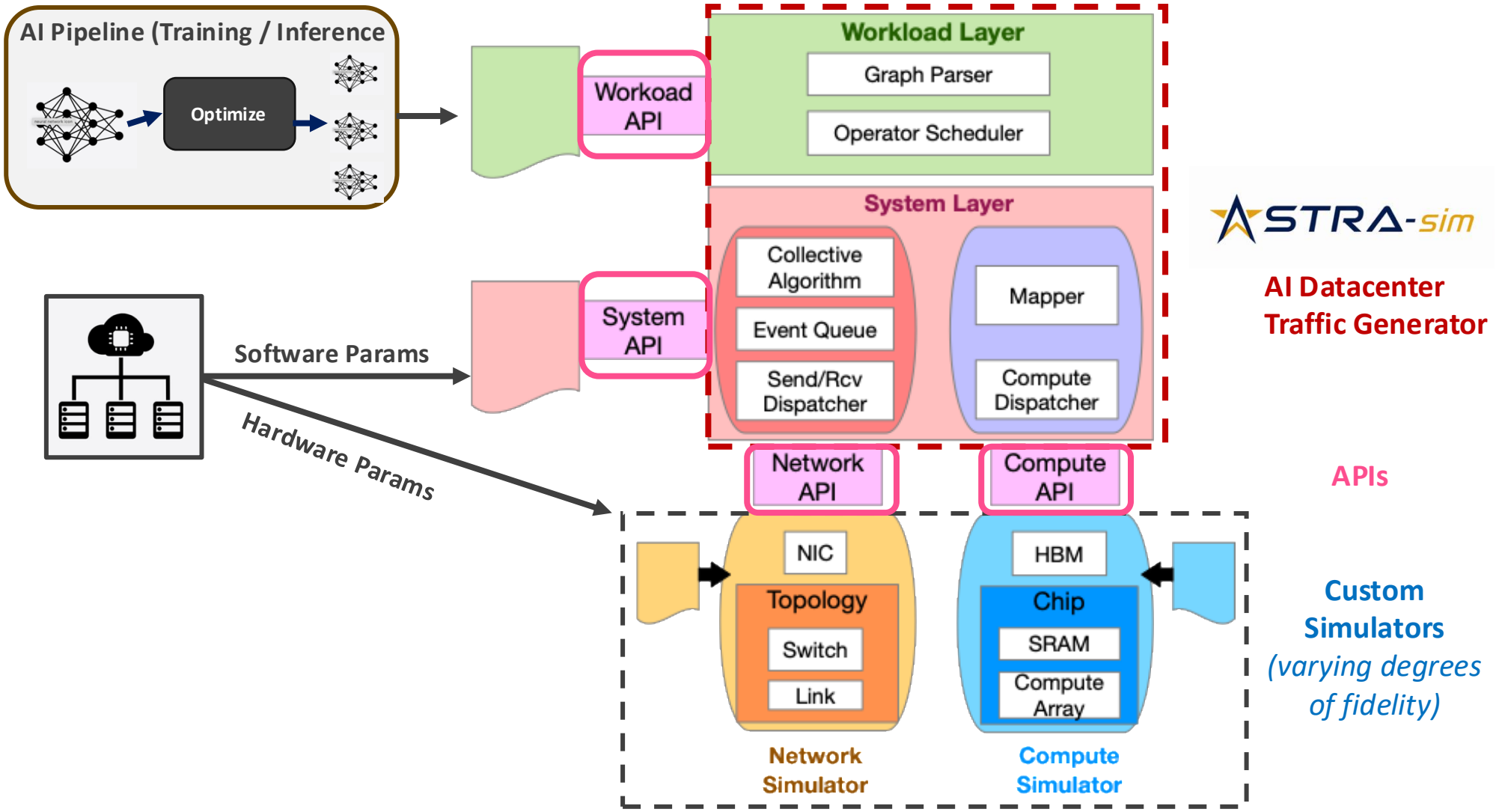
Figure Courtesy: Venkat Pullela (Keysight)

Challenge: Cross-Coupled HW-SW Co-Design-Space



Overview of the ASTRA-sim Ecosystem

https://astra-sim.github.io/



ASTRA-sim

AI Datacenter Traffic Generator

APIs

Custom Simulators *(varying degrees of fidelity)*

- AI Model Architecture
- Parallelization Strategy
- Communication Policy
- Framework-level Scheduling
- Collective Scheduling
- Compiler Optimizations
- Transport Layer
- HW specific Optimizations
- Scale-out Topology
- Memory & Storage Hierarchy
- Scale-up/in Topology
- Dataflow
- Flow Control
- Microarch
- End Point
- Precision
- Integration Technology
- Link Technology
- Memory Technology

The ASTRA-sim Journey

(2020)

ASTRA-SIM: Enabling SW/HW Co-Design Exploration for Distributed DL Training Platforms

Saeed Rashidi*, Srinivas Sridharan†, Sudarshan Srinivasan‡ and Tushar Krishna*

(2023)

ASTRA-sim2.0: Modeling Hierarchical Networks and Disaggregated Systems for Large-model Training at Scale

William Won*§, Taekyung Heo*§, Saeed Rashidi*§, Srinivas Sridharan†, Sudarshan Srinivasan‡, Tushar Krishna*

(2026)

ASTRA-sim 3.0: Next-Level Distributed Machine Learning Simulations via High-Fidelity GPU and Infrastructure Modeling

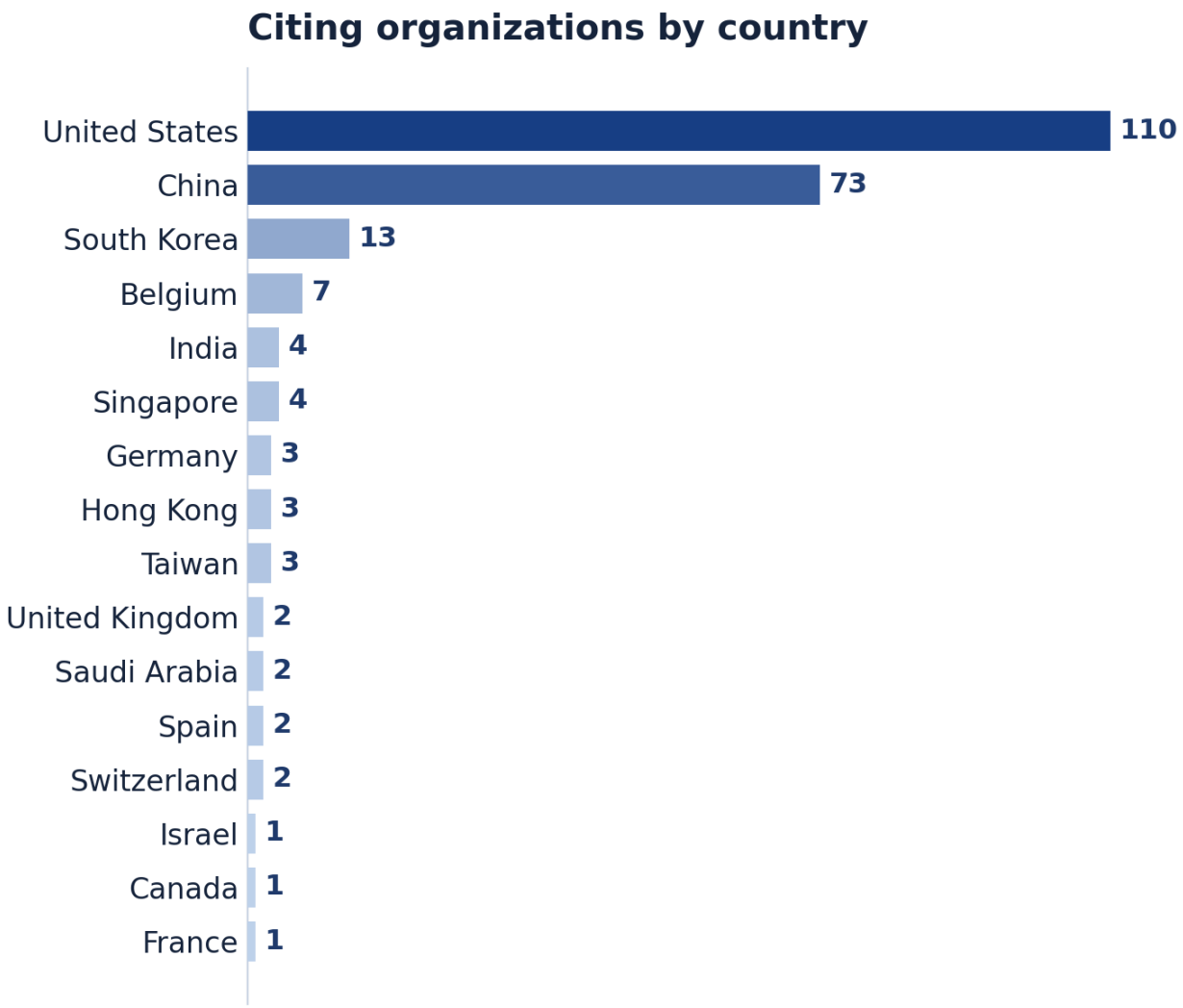
William Won*[1], Jinsun Yoo*[2], Tuan Ta*[1], Moumita Dey*[1], Andy Balogh[3], Pradosh Datta[3], Furkan Eris[1], Conor Green[1 4], Winston Liu[3], Changhai Man[2], Kingshuk Mandal[3], Amos Rai[3], Vinay Ramakrishnaiah[1], Ruchi Shah[1], David Sidler[1], Harsh Sikhwal[3], Hanjiang Wu[2], Tushar Krishna†[2], and Bradford M. Beckmann†[1]

MLCommons Chakra: Advancing Performance Benchmarking and Co-design using Standardized Execution Traces

Srinivas Sridharan[1] Andy Balogh[2] Bradford M. Beckmann[3] Brian Coutinho[1] Louis Feng[4] Sheng Fu[1] Sanshan Gao[1] Mehryar Garakani[5] Taekyung Heo[1] David Kanter[6] Josh Ladd[1] Ziwei Li[7] Winston Liu[2] Changhai Man[7] Dan Mihailescu[2] Spandan More[3] Joongun Park[7] Ashwin Ramachandran[4] Vinay Ramakrishnaiah[3] Saeed Rashidi[4] Vijay Janapa Reddi[8] Puneet Sharma[9] Phio Tian[1] William Won[3 7] Hanjiang Wu[7] Huan Xu[7] Jinsun Yoo[7] Tushar Krishna[7 10]

https://github.com/mlcommons/chakra

Worldwide Research Adoption Today



Sources: Semantic Scholar + OpenAlex; GitHub API. As of Jun 2026.

GITHUB · astra-sim/astra-sim

open-sourced 2020 · 15 watchers

★ 617 STARS

210 FORKS

192 citing papers

71 use it in evaluation

105 distinct institutions

16 countries

WHO USES IT — BY SECTOR

Universities	76
Industry	18
Research / gov labs	11

Resources

- **Main Website:** https://astra-sim.github.io/
- **Tutorial Website:** https://astra-sim.github.io/tutorials/isca-2026
- **Past Tutorials*:** https://astra-sim.github.io/tutorials
- **Codebase:** https://astra-sim.github.io/tutorials
- **Wiki:** https://astra-sim.github.io/astra-sim-docs/index.html
- **Papers:**
  - **ASTRA-sim3.0:** https://arxiv.org/abs/2606.10440
  - **MLCommons Chakra:** https://arxiv.org/abs/2605.11333
  - **ASTRA-sim2.0:** https://arxiv.org/abs/2303.14006
  - **ASTRA-sim1.0:** https://sites.gatech.edu/ece-synergy/files/2020/08/astrasim_ispass2020.pdf

*Slide Content for this tutorial adapted from past tutorials
(Acknowledgments: Saeed Rashidi, Taekyung Heo, Joongun Park)

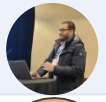
Morning Agenda

All times EDT

Time	Title	Presenter
8:00 am	Introduction	 Tushar Krishna — Georgia Tech
8:15 am	Workload and System layer	 Will Won — AMD
8:35 am	Chakra Trace Generation	 Changhai Man — Georgia Tech
8:50 am	Network Layer and ns-3	 Jinsun Yoo — Georgia Tech
9:10 am	Extending ASTRA-sim with HTSim for Ultra Ethernet Simulation	 Veerasenareddy Burru — Marvell
9:30 am	InfraGraph: Vendor-Neutral Infrastructure Topology for AI/HPC	 Harsh Sikhwal — Keysight
10:00 am	Coffee Break	
10:30 am	Customized Collectives Algorithms with MSCCLPP	 Ruchi Shah — AMD
11:00 am	Detailed GPU Model in System Layer	 Tuan Ta — AMD
11:30 am	Inter-GPU Communication Protocols & Synchronization	 Moumita Dey — AMD

Afternoon Agenda

All times EDT

Time	Title	Presenter
1:40 pm	Lessons from a Quarter-Century of Computer Architecture Simulation	 Brad Beckmann — AMD
2:10 pm	Beyond Communication Modeling: Extending ASTRA-sim for End-to-End AI Infrastructure Design and Evaluation	 Puneet Sharma — HPE
2:30 pm	ASTRA-sim-service: A Composable Simulation Interface for AI Infrastructure Co-Design Across Workloads and Fidelity Tiers	 Harsh Sikhwal — Keysight
2:50 pm	Hespas: Multi-Fidelity Simulation of StableHLO-ML Workloads with ASTRA-sim	 Abubakr Nada — imec
3:10 pm	A Co-Design Framework for Heterogeneous Multi-Stage LLM Inference using ASTRA-SIM	 Abhimanyu Bambhaniya — Georgia Tech / InfraVana
3:30 pm	Coffee Break	
4:00 pm	Enabling Memory Tiering in ASTRA-sim with Extended Chakra Traces	   Ulf Hanbutte & Nikhil Stephen & Shuting Du — Marvell
4:20 pm	Enabling Energy Efficiency for Distributed Machine Learning	  Tanvir Khan & Tawhid Bhuiyan — Columbia University
4:40 pm	Modeling Multi-tenant Scheduling in ML Clusters using Astra-sim	 Shawn Chen — CMU