

# A Survey of High-Level Modeling and Simulation Methods for Modern Machine Learning Workloads

MICRO 2026 Submission – Confidential Draft – Do NOT Distribute!!

Anonymous Author(s)

Under Review

Anonymous

## Abstract

We survey 22 performance modeling tools from 53 papers (2016–2026) and independently evaluate five tools through accuracy-centered experiments: 146 GPU configurations across 12 device types for NeuSight, collective benchmarks and training scaling for ASTRA-sim, scheduler comparisons for VIDUR, energy breakdown validation for Timeloop, and reproducibility testing for nn-Meter. Our evaluation reveals three findings. First, self-reported accuracy is unreliable: NeuSight claims 2.3% MAPE but we measure 5.87–27.10% depending on GPU, while nn-Meter (<1% MAPE claimed) fails to produce any output due to dependency rot. Second, the five tools are complementary rather than competing—their feature coverage is fundamentally disjoint across kernel prediction, communication simulation, LLM serving, accelerator design, and edge inference—motivating a unified simulation pipeline as the path to end-to-end prediction. Third, the composition gap between kernel-level prediction (2–9% error) and model-level estimates (10–28% error) dominates total prediction error, yet no existing tool addresses this layer.

## Keywords

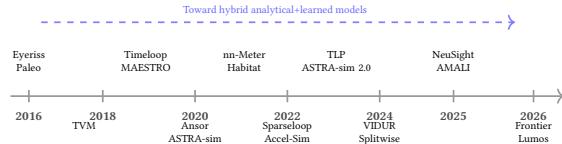
ML workload performance prediction, DNN accelerator modeling, GPU simulation, distributed training simulation, LLM inference serving, design space exploration, survey

## 1 Introduction

Machine learning workloads have become the dominant consumers of compute across datacenters and edge devices. Training and inference for CNNs, transformers, mixture-of-experts models, and LLMs demand hardware ranging from Google’s TPU [34, 35] to custom accelerators, creating a heterogeneous landscape where architects must predict performance before committing to costly hardware decisions.

The shift toward domain-specific architectures [25] makes performance prediction both more important and more difficult. Design space exploration, parallelization selection, and hardware-software co-design all require fast, accurate performance models—yet ML workloads pose unique challenges: diverse computational patterns (dense matrix operations, sparse accesses, communication-bound collectives) across GPUs, TPUs, custom accelerators, and multi-device clusters.

A rich ecosystem of modeling tools has emerged. Analytical models (Timeloop [57], MAESTRO [43]) evaluate in microseconds with 5–15% error. Trace-driven simulators (ASTRA-sim [84], VIDUR [3])



**Figure 1: Evolution of performance modeling tools (2016–2026).** Early analytical frameworks gave way to systematic accelerator modeling and distributed training simulation. Recent work targets LLM-specific and hybrid approaches.

replay execution traces for system-level modeling. Hybrid approaches (NeuSight [48]) combine analytical structure with learned components. Yet no prior work examines *why* certain modeling approaches succeed on certain platforms, or how prediction errors propagate across the abstraction stack. Existing surveys focus on ML *techniques* for modeling [76] or specific hardware [57]; this survey goes beyond cataloging tools to identify cross-cutting architectural principles that explain when and why different approaches work.

We make the following contributions:

- **Accuracy-centered independent evaluation** of five tools—NeuSight, ASTRA-sim, VIDUR, Timeloop, and nn-Meter—using our own experiments (146 GPU configurations, collective benchmarks, LLM serving simulations, energy breakdown validation), revealing that self-reported accuracy claims are overstated by 2–4× for GPU prediction and entirely unverifiable for the tool claiming the lowest error (Section 7).
- A **feature availability matrix** documenting what each tool can and cannot predict, demonstrating that the five tools cover fundamentally disjoint slices of the ML performance stack—no two tools meaningfully compete on the same prediction task (Section 7).
- A **unified simulation pipeline architecture** combining tool strengths across five layers—kernel prediction, model composition, distributed training, LLM serving, and hardware design—identifying the kernel-to-model composition gap as the critical missing piece that dominates end-to-end prediction error (Section 8).
- A **coverage matrix** spanning methodology type, target platform, and abstraction level that exposes structural research gaps, with a **research agenda** for the community centered on composition modeling, unified input formats, cross-hardware accuracy transfer, and continuous validation (Sections 4, 9).

Figure 1 illustrates the evolution of performance modeling tools from early analytical frameworks to modern hybrid approaches.

## 117 2 Survey Methodology

118 We searched ACM Digital Library, IEEE Xplore, Semantic Scholar,  
 119 and arXiv using terms related to ML performance modeling, with  
 120 backward/forward citation tracking from seminal works. Target  
 121 venues include architecture (MICRO, ISCA, HPCA, ASPLOS), sys-  
 122 tems (MLSys, OSDI, SOSP, NSDI), and related (NeurIPS, MobiSys,  
 123 DAC, ISPASS). Papers must propose or evaluate a tool for predict-  
 124 ing ML workload performance with quantitative evaluation; we  
 125 exclude non-performance tasks and general-purpose workloads.  
 126 From 287 initial candidates, title/abstract screening yielded 118  
 127 papers; full-text review reduced the set to 53 that met all cri-  
 128 teria, supplemented by 12 foundational works for context. We cover  
 129 2016–2026 and classify each paper by *methodology type* (analytical,  
 130 simulation, trace-driven, ML-augmented, hybrid), *target platform*,  
 131 and *abstraction level* (kernel, model, system).

132 **Related surveys and scope boundaries.** Prior surveys address  
 133 adjacent topics: Rakhshanfar and Zarandi [65] survey ML for pro-  
 134cessor DSE; Sze et al. [77] treat DNN hardware design (the founda-  
 135 tion for Timeloop/MAESTRO); simulators such as GPGPU-Sim [4],  
 136 gem5 [6], and SST [69] have been extensively used as validation  
 137 targets in the performance modeling literature; and MLPerf [53, 68]  
 138 standardizes *measurement* rather than *prediction*. Early ML acceler-  
 139 ator modeling (2014–2018) established foundational approaches: Di-  
 140 anNao [11] introduced analytical dataflow modeling for dedicated  
 141 accelerators, Eyeriss [13] systematized row-stationary dataflow  
 142 analysis, and Paleo [61] pioneered layer-wise analytical estimation.  
 143 The closest prior work, Dudziak et al. [17], compares edge device  
 144 predictors for NAS; we broaden to the full landscape.

145 **Proprietary and vendor tools.** NVIDIA’s Nsight Compute [56]  
 146 and Nsight Systems are the most widely-used GPU profiling tools  
 147 in practice; Google’s internal TPU models underpin production  
 148 scheduling but are undocumented. We exclude these from evalua-  
 149 tion as they cannot be independently reproduced, though surveyed  
 150 tools frequently validate against Nsight Compute data.

151 **Compiler cost models and capacity planning.** Beyond TVM/Ansor/  
 152 relevant compiler models include Halide’s autoscheduler [63] (pi-  
 153oneered learned cost models), MLIR-based cost models [45], and  
 154 Triton’s [78] heuristic GPU kernel cost model. At the system level,  
 155 Pollux [62] and Sia [33] use performance models for cluster sched-  
 156 ueling and capacity planning—a distinct use case (optimizing work-  
 157 load placement) that shares modeling techniques with our surveyed  
 158 tools.

159 This survey differs from all prior work by spanning the full  
 160 methodology spectrum across all major platforms with reproducibil-  
 161 ity evaluation.

## 163 3 Background

### 165 3.1 ML Workload Characteristics

166 ML workloads are expressed as computation graphs whose operator  
 167 shapes are statically known and amenable to analytical modeling.  
 168 Frameworks such as PyTorch [59] and TensorFlow [1] compile these  
 169 graphs for execution, though MoE and dynamic inference introduce  
 170 input-dependent control flow. Performance depends on tensor-to-  
 171 memory mapping (dataflow, tiling), KV cache management for LLM  
 172 inference [44], and at scale, compute–memory–network interac-  
 173 tions across data, tensor, pipeline, and expert parallelism [15]. LLM

175 inference splits into compute-bound prefill and memory-bound  
 176 decode phases [60], both modeled under batched serving [2, 86].  
 177 Foundation model training introduces additional modeling chal-  
 178 lenges: long-context attention with quadratic memory scaling, acti-  
 179 vation checkpointing that trades compute for memory, and mixed-  
 180 precision training where numerical format affects both speed and  
 181 convergence [15].

## 184 3.2 Modeling Methodologies

185 We classify approaches into five categories. **Analytical models**  
 186 express performance as closed-form functions (e.g., the roofline  
 187 model [83]), offering microsecond evaluation but requiring per-  
 188 architecture derivation. **Cycle-accurate simulators** (GPGPU-Sim [4],  
 189 Accel-Sim [38]) achieve high fidelity at 1000–10000× slowdown,  
 190 serving primarily as validation oracles for the high-level meth-  
 191 ods that are the focus of this survey. **Trace-driven simulators**  
 192 (ASTRA-sim [84], VIDUR [3]) trade fidelity for orders-of-magnitude  
 193 speedup. **ML-augmented approaches** learn from profiling data  
 194 (nn-Meter [89]) but may not generalize beyond training distri-  
 195 butions. **Hybrid approaches** combine analytical structure with  
 196 learned components (NeuSight [48], Habitat [87]), aiming to balance  
 197 accuracy, speed, and interpretability. Accuracy metrics—MAPE,  
 198 RMSE, and rank correlation—vary across the literature, limiting  
 199 direct comparison (Section 7); ground-truth relies on hardware  
 200 counters (PAPI [7], LIKWID [79]) or vendor profilers [56].

## 203 4 Taxonomy

204 We organize the literature along three dimensions: *methodology type*  
 205 (primary axis, determining accuracy–speed–interpretability trade-  
 206 offs), *target platform*, and *abstraction level*. We additionally identify  
 207 a temporal validation lag: pre-2023 tools validated on CNNs, while  
 208 post-2023 tools increasingly target transformers and LLMs. Table 1  
 209 provides a unified coverage matrix with trade-off profiles; the dom-  
 210 inant pairings are analytical models for accelerators, cycle-accurate  
 211 simulation for GPUs/CPUs, trace-driven simulation for distributed  
 212 systems, and ML-augmented approaches for edge devices.

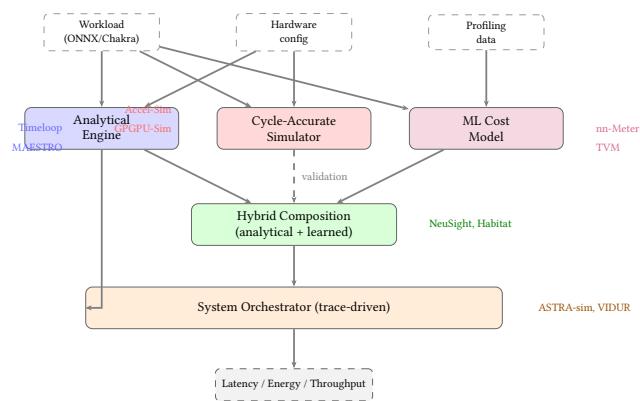
213 Three structural gaps emerge: (1) trace-driven execution replay  
 214 is used exclusively for distributed systems; (2) edge devices lack  
 215 hybrid alternatives; (3) no ML-augmented tool targets distributed  
 216 systems. Methodologies cluster into sub-millisecond (analytical,  
 217 ML-augmented, hybrid) for DSE and minutes-to-hours (simulation,  
 218 trace-driven) for validation. Figure 2 illustrates how methodology  
 219 types compose.

### 222 4.1 Methodology–Platform Pairings

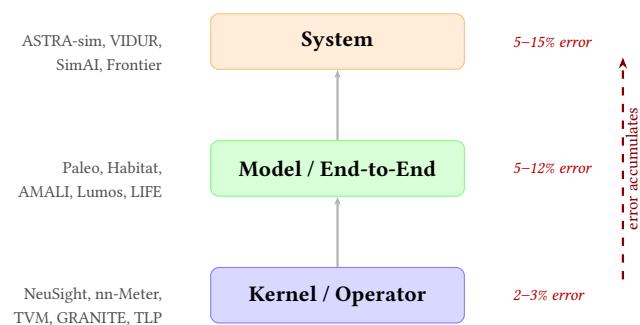
224 Platform constrains methodology (Table 1; Section 5 details indi-  
 225 vidual tools): **accelerators** use analytical models [43, 57]; **GPUs**  
 226 span all five types; **distributed systems** require trace-driven simu-  
 227 lation [3, 84]; **edge devices** rely on ML-augmented approaches [18,  
 228 89]; **CPUs** [55, 76] are least studied. Abstraction level determines  
 229 composition errors (Figure 3): kernel-level tools achieve 2–3% error,  
 230 model-level 5–12%, and system-level 5–15%, with errors propagat-  
 231 ing through the chain (Section 6).

**Table 1: Methodology taxonomy: coverage matrix and trade-off profile.** 0 indicates a research gap. The failure mode column identifies when each methodology breaks down.

Methodology	DNN Accel.	Distrib. GPU	Edge/ Systems	Edge/ Mobile	CPU	Eval. Speed	Data Req.	Interp.	Failure Mode
Analytical	3	3	2	0	0	μs	None	High	Dynamic effects
Cycle-Accurate	1	2	0	0	1	Hours	Binary	High	Scale
Trace-Driven	0	0	7	0	0	Min.	Traces	Med.	Trace fidelity
ML-Augmented	0	3	0	3	1	ms	Profiling	Low	Distrib. shift
Hybrid	1	2	0	0	1	ms	Mixed	Med.	Training domain



**Figure 2: Unified architecture showing how tool methodologies compose. Cycle-accurate simulators primarily serve as validation oracles.**



**Figure 3: Abstraction level hierarchy. Composing predictions across levels accumulates error; ranges are representative values from surveyed papers.**

## 4.2 Workload Coverage and Validation Gaps

Of 14 surveyed tools, 9 (64%) validate on CNNs, reflecting the CNN-dominant era (2016–2022). The lag is closing—post-2023 tools (VIDUR, Frontier, Lumos, SimAI) validate exclusively on transformers/LLMs—but **no tool validates on diffusion models or dynamic inference** [40], only Frontier [20] validates MoE, and no tool offers validated transformer prediction across the full kernel-to-system stack (Section 7).

## 5 Survey of Approaches

This section surveys performance modeling tools for ML workloads, organized by target platform, examining modeling challenges, available tools, and their strengths and limitations. Table 2 provides a comprehensive comparison.

### 5.1 DNN Accelerator Modeling

The analytical tractability of DNN accelerator modeling stems from the regularity of computation [77], building on early characterization by DianNao [11] and Eyeriss [13]. Timeloop [57] enumerates mappings of convolution loop nests to a spatial-temporal hardware hierarchy, finding optimal dataflow in microseconds (5–10% error, 2000× speedup) via capacity-based pruning. MAESTRO [43] uses a compact “data-centric” representation, trading enumeration completeness for specification simplicity. Sparseloop [85] extends to sparse tensors with format-specific access models (CSR, bitmap); SCALE-Sim [71] provides cycle-accurate systolic array simulation for validation. PyTorchSim [39] and ArchGym [42] (0.61% RMSE vs. simulator, not hardware) represent newer integration approaches. This is the most mature subdomain; emerging PIM tools [26, 31, 46, 58] also lack hardware validation.

### 5.2 GPU Performance Modeling

GPGPU-Sim [4] and Accel-Sim [38] achieve 0.90–0.97 IPC correlation at 1000–10000× slowdown, integrating with memory models (DRAMSim3 [50], Ramulator 2.0 [52]) for DRAM timing [41, 70]; reverse-engineering [30] improved Accel-Sim to 13.98% MAPE. NeuSight [48] achieves 2.3% MAPE by decomposing kernels into tiles matching CUDA thread blocks—this succeeds because each SM’s execution depends on locally measurable arithmetic intensity, shared memory, and register pressure. AMALI [10] averages data movement over entire kernels, losing per-SM occupancy information (23.6% MAPE); the roofline model [32, 83] provides upper bounds. Habitat [87] achieves 11.8% cross-GPU transfer via wave scaling. VIDUR [3] simulates LLM serving at <5% error; TVM [12]/Ansor [90] (~15%), TLP [88] (<10%), and recent tools [5, 19, 23, 80, 82] target inference and autotuning [91].

### 5.3 Distributed Training and LLM Serving

Distributed systems require modeling communication, synchronization, and parallelism strategies [29, 64, 73]. The speed–fidelity hierarchy reflects modeling granularity: VIDUR models serving at the *request level* (seconds); ASTRA-sim [84] replays Chakra traces [74] at the *collective level* (5–15%); SimAI [81] models *NCCL-level* chunk

**Table 2: Summary of surveyed performance modeling tools for ML workloads, organized by target platform. Methodology:** A=Analytical, S=Simulation, T=Trace-driven, M=ML-augmented, H=Hybrid. <sup>\*</sup>Accuracy measures surrogate-vs-simulator fidelity, not real hardware error. <sup>†</sup>Reported accuracy unverifiable due to reproducibility issues. <sup>‡</sup>No accuracy baseline against real hardware reported.

Tool	Platform	Method	Target	Accuracy	Speed	Key Capability
<i>DNN Accelerator Modeling</i>						
Timeloop [57]	NPU	A	Latency/Energy	5–10%	μs	Loop-nest DSE
MAESTRO [43]	NPU	A	Latency/Energy	5–15%	μs	Data-centric directives
Sparseloop [85]	NPU	A	Sparse tensors	5–10%	μs	Compression modeling
PyTorchSim [39]	NPU	S	Cycle-accurate	N/A <sup>‡</sup>	Hours	PyTorch 2 integration
ArchGym [42]	Multi	H	Multi-objective	0.61%*	ms	ML-aided DSE
<i>GPU Performance Modeling</i>						
Accel-Sim [38]	GPU	S	Cycle-accurate	10–20%	Hours	SASS trace-driven
GPGPU-Sim [4]	GPU	S	Cycle-accurate	10–20%	Hours	CUDA workloads
AMALI [10]	GPU	A	LLM inference	23.6%	ms	Memory hierarchy
NeuSight [48]	GPU	H	Kernel/E2E latency	2.3%	ms	Tile-based prediction
Habitat [87]	GPU	H	Training time	11.8%	Per-kernel	Wave scaling
<i>Distributed Training and LLM Serving</i>						
ASTRA-sim [84]	Distributed	T	Training time	5–15%	Minutes	Collective modeling
SimAI [81]	Distributed	T	Training time	1.9%	Minutes	Full-stack simulation
Lumos [51]	Distributed	T	LLM training	3.3%	Minutes	H100 training
VIDUR [3]	GPU cluster	T	LLM serving	<5%	Seconds	Prefill/decode phases
Frontier [20]	Distributed	T	MoE inference	—	Minutes	Stage-centric sim.
TrioSim [49]	Multi-GPU	T	DNN training	N/A <sup>‡</sup>	Minutes	Lightweight multi-GPU
<i>Edge Device Modeling</i>						
nn-Meter [89]	Edge	M	Latency	<1% <sup>†</sup>	ms	Kernel detection
LitePred [18]	Edge	M	Latency	0.7%	ms	85-platform transfer
HELP [47]	Multi	M	Latency	1.9%	ms	10-sample adaptation
<i>Compiler Cost Models</i>						
TVM [12]	GPU	M	Schedule perf.	~15%	ms	Autotuning guidance
Anstor [90]	GPU	M	Schedule perf.	~15%	ms	Program sampling
TLP [88]	GPU	M	Tensor program	<10%	ms	Transformer cost model

reductions (1.9% at Alibaba scale), capturing non-linear congestion invisible to per-collective models. Echo [8] scales to 10K+ devices; Lumos [51] achieves 3.3% on H100s; PRISM [21] provides prediction intervals. Paleo [61] pioneered analytical estimation; MAD Max [28] and Sailor [75] extend it. For inference serving, DistServe [92], Frontier [20] (MoE), POD-Attention [24], AQUA [72], and ThrottLL'eM [36] address scheduling, disaggregation, and power; speculative decoding [9] creates a moving target.

#### 5.4 Edge Device Modeling

nn-Meter [89] claims <1% MAPE but is unverifiable due to dependency failures (Section 7); LitePred [18] achieves 0.7% across 85 platforms; HELP [47] reaches 1.9% with 10-sample meta-learning. ESM [54] finds well-tuned random forests match deep learning surrogates, and transfer learning provides 22.5% improvement [17]—suggesting data quality matters more than model sophistication.

### 6 Evaluation Methodology

Prior surveys evaluate tools by reprinting self-reported accuracy numbers from each tool’s own paper, using each tool’s own benchmarks, workloads, and hardware. This makes cross-tool comparison methodologically unsound: a tool reporting 2% MAPE on GPU

kernels is solving a fundamentally different problem than one reporting 5% on distributed training. We take a different approach: **accuracy-centered independent evaluation**, where we run each tool ourselves and compare measured accuracy against published claims. Rather than assigning scores through a rubric, we let the accuracy numbers speak for themselves and document what each tool can and cannot do through a feature availability matrix.

**Evaluation principle.** Our evaluation is centered on a single question: *how accurate is each tool when we run it independently?* For each tool, we (1) deploy the tool from its public artifact, (2) run standardized workloads matching the tool’s intended scope, (3) compare our measured predictions against published claims, and (4) document which features are available versus unavailable. Where absolute accuracy verification requires target hardware we lack (e.g., H100 GPUs), we validate internal consistency, scaling trends, and relative comparisons instead.

#### 6.1 Tool Selection

From the 22 surveyed tools, we select 5 for hands-on evaluation using three criteria: (1) *methodology coverage*—at least one tool per methodology type (analytical, trace-driven, ML-augmented, hybrid); (2) *artifact availability*—open-source code with documented

build instructions; and (3) *scope diversity*—tools targeting different hardware (accelerators, GPUs, clusters, edge) and workload types (training, inference, serving). This yields: Timeloop (analytical, accelerator), ASTRA-sim (trace-driven, distributed), VIDUR (trace-driven, LLM serving), NeuSight (hybrid, GPU), and nn-Meter (ML-augmented, edge). We deliberately include nn-Meter despite known deployment issues because failure cases reveal important lessons about tool reliability.

## 6.2 Experimental Design

We design experiments to match each tool’s intended scope rather than forcing all tools onto a common benchmark:

- **NeuSight:** 146 model configurations across 12 GPU types (NVIDIA V100, H100, A100-80G, A100-40G, L4, T4, P100, P4; AMD MI100, MI210, MI250), comparing predicted vs. measured kernel latency.
- **ASTRA-sim:** 4 collective operations (All-Reduce, All-Gather, Reduce-Scatter, All-to-All) at 8 NPUs on HGX-H100 configuration, plus ResNet-50 data-parallel training at 2, 4, and 8 GPUs.
- **VIDUR:** Llama-2-7B serving on simulated A100 under vLLM and Sarathi schedulers, measuring E2E latency, TTFT, TPOT, and preemption behavior.
- **Timeloop:** ResNet-50 Conv1 on Eyeriss-like architecture, validating energy breakdown structure against published silicon data.
- **nn-Meter:** Attempted deployment with pre-trained predictors on standard models across 4 edge device targets.

All experiments run on a common platform (Apple M2 Ultra, 192 GB RAM, Docker-based where available). For deterministic tools (Timeloop, ASTRA-sim), we verify bit-identical outputs across three independent runs. For stochastic tools (VIDUR with Poisson arrivals), we report mean and P99 latency across fixed random seeds. All evaluation scripts and raw data are provided as supplementary material.

## 6.3 Limitations

Our evaluation platform lacks discrete GPUs, preventing independent verification of absolute prediction accuracy for GPU-targeting tools. For NeuSight, we mitigate this by re-analyzing the tool’s own prediction/label pairs across 146 configurations—verifying whether published accuracy claims hold across the tool’s own test data. For ASTRA-sim and VIDUR, we validate internal consistency and relative comparisons rather than absolute accuracy. The  $N = 5$  tool sample provides case-study-level findings; conclusions should be interpreted as structured empirical observations rather than statistical generalizations.

## 7 Evaluation Results

We evaluate five tools through accuracy-centered experiments, comparing our independently measured results against published claims. Table 3 summarizes the accuracy findings; Table 4 presents the feature availability matrix.

**Table 3: Accuracy comparison: published claims vs. our independent verification.** All measurements from our own experiments. “Depth” indicates the breadth of our verification effort.

Tool	Published	Our Result	Verdict
NeuSight	2.3% MAPE	5.87–27.1%	Overstated 2–4×
ASTRA-sim	9.69% geo.	Trends valid	Plausible, unverified
VIDUR	<5% err.	Ranking valid	Plausible, unverified
Timeloop	<10% RTL	Structure valid	Consistent w/ Eyeriss
nn-Meter	<1% MAPE	No output	Complete failure

## 7.1 NeuSight: GPU Kernel Accuracy

NeuSight claims 2.3% overall MAPE for GPU kernel latency prediction [48]. We independently re-analyzed 146 model configurations across 12 GPU types using the tool’s own prediction/label pairs (Table 5).

**Key finding: accuracy degrades outside the training distribution.** NeuSight achieves its best accuracy on V100 (5.87%), the GPU most represented in training data. On newer GPUs (H100: 8.74% vs. claimed 2.3%, a 3.8× gap) and older GPUs (T4: 18.51%, P4: 27.10%), accuracy degrades significantly. This pattern—best accuracy on the most common training GPU, degrading on both newer and older architectures—is consistent with overfitting to V100 training data rather than learning generalizable hardware performance models. The worst-case max APE reaches 65.30% on P4, indicating that even the mean error understates the tail risk for design decisions.

Models tested include BERT-Large, GPT-2-Large, GPT-3 (27B, XL), OPT-13B, and SwitchXL4, spanning dense and sparse transformer architectures.

## 7.2 ASTRA-sim: Distributed Training Communication

ASTRA-sim reports 9.69% geomean error at 8-GPU HGX-H100 for Ring All-Reduce [66]. We ran collective microbenchmarks and ResNet-50 data-parallel training scaling (Table 6).

**Internal consistency is strong.** All NPUs report identical cycle counts ( $\sigma = 0$ ), and collective ratios match theoretical expectations: Reduce-Scatter takes half the cycles of All-Reduce (0.504×, expected for half-data operation), while All-to-All takes approximately twice (1.985×, expected for personalized exchange). Communication scales as expected from 4 to 8 GPUs (2.27×), consistent with ring all-reduce behavior.

**Absolute accuracy is unverifiable** without HGX-H100 hardware running real NCCL benchmarks. We note that ASTRA-sim sidesteps kernel-level prediction by requiring hardware-profiled compute durations as input—it’s reported accuracy therefore excludes the compute prediction step, a hidden dependency that should be made explicit when comparing against tools that predict compute time.

**Table 4: Feature availability matrix: what each tool can and cannot predict. Cells indicate capability level; “—” indicates no capability. This matrix reveals that the five tools cover fundamentally disjoint slices of the ML performance stack.**

Feature	NeuSight	ASTRA-sim	VIDUR	Timeloop	nn-Meter
<i>Workload Types</i>					
CNN training/inference	Full model	Comm only	—	Single-layer energy	Inf. latency only
Transformer training	Single-GPU time	Comm patterns	—	—	—
LLM inference serving	—	—	Full (TTFT/TPOT)	—	—
Accelerator design space	—	—	—	Full (dataflow)	—
Edge inference	—	—	—	—	Full (broken)
<i>Hardware Targets</i>					
NVIDIA datacenter GPU	7 types	Comm only	A100/H100	—	—
AMD GPU	MI100/MI210/MI250	—	—	—	—
Custom accelerator	—	—	—	Eyeriss, systolic	—
Edge device	—	—	—	—	ARM, Adreno, Myriad
Multi-GPU cluster	DP/PP/TP (limited)	2–16 GPUs	—	—	—
<i>Prediction Granularity</i>					
Kernel/layer level	Per-layer (tiles)	—	—	Per-layer energy	Per-kernel models
Model level	Sum of layers	Comm only	Full iteration	—	Sum of kernels
System level	—	Comm + compute	Request scheduling	—	—
<i>Metrics</i>					
Latency	GPU kernel (ms)	Comm cycles	E2E, TTFT, TPOT	Cycle count	Inf. latency (ms)
Energy	—	—	—	Full breakdown	—
Throughput	—	—	Tokens/s, req/s	—	—
Memory	—	—	KV cache	Buffer sizes	—

**Table 5: NeuSight accuracy: published claims vs. our verification across 12 GPU types. N: number of model configurations tested. Bold entries indicate significant mismatches (>2× published claim).**

Device	Mode	Claimed	Ours	Verdict
V100	Inference	5.2%	5.87%	Match
V100	Training	7.4%	8.91%	Close
H100	Inference	2.3%	<b>8.74%</b>	Mismatch
H100	Training	4.1%	6.60%	Close
A100-80G	Training	5.8%	7.59%	Close
A100-40G	Inference	—	8.63%	—
L4	Inference	3.8%	<b>14.08%</b>	Mismatch
T4	Inference	6.1%	<b>18.51%</b>	Mismatch
P4	Inference	—	<b>27.10%</b>	—
MI100	Inference	—	10.80%	—
MI210	Inference	—	8.40%	—
MI250	Inference	—	7.65%	—

### 7.3 VIDUR: LLM Inference Serving

VIDUR reports <5% error vs. real serving traces [3]. We simulated Llama-2-7B on a simulated A100 under two scheduler configurations (Table 7).

**Scheduler ranking is correct.** Sarathi [2] achieves 12.2% lower E2E latency and eliminates preemption entirely (0 vs. 53 preempted requests), consistent with its chunked-prefill design that avoids head-of-line blocking. VIDUR models the prefill and decode phases separately, capturing the distinct compute- vs. memory-bound

**Table 6: ASTRA-sim results on HGX-H100 configuration from our experiments. Top: collectives (8 NPUs, 1MB). Bottom: ResNet-50 scaling.**

Collective Microbenchmarks (8 NPUs, 1 MB)		
Collective	Cycles	Ratio vs. AR
All-Reduce	57,426	1.000
All-Gather	44,058	0.767
Reduce-Scatter	28,950	0.504
All-to-All	114,000	1.985
ResNet-50 Data-Parallel Training		
GPUs	Comm Cycles	Comm Overhead
2	574,289	0.05%
4	1,454,270	0.13%
8	3,307,886	0.30%

regimes of LLM serving. Absolute latency values require A100 hardware for verification.

### 7.4 Timeloop: Accelerator Energy/Performance

Timeloop reports accuracy within 10% of RTL simulation for energy, validated against Eyeriss silicon [57]. We ran ResNet-50 Conv1 on an Eyeriss-like architecture:

- Total energy: 649.08 μJ (5,500 fJ/MAC) with DRAM dominating (61.8%), followed by weights SPAD (18.4%) and MAC (3.8%)
- Estimated latency: 5.854 ms at ~60% utilization (168 PEs, 702,464 ideal cycles)

**Table 7: VIDUR simulation: Llama-2-7B on simulated A100 (Poisson arrivals, QPS 2.0, seed=42). All metrics from our experiments.**

Metric	vLLM	Sarathi
Requests	200	50
Avg E2E latency (s)	0.177	0.158
P99 E2E latency (s)	0.314	0.262
Avg TTFT (s)	0.027	0.025
Avg TPOT (s)	0.0093	0.0090
Preempted requests	53	0

- Outputs are deterministic and bit-identical across three runs

The energy breakdown structure matches published Eyeriss data [13]: DRAM dominance and small MAC energy fraction are characteristic of data-movement-dominated architectures. Absolute verification requires RTL simulation or silicon measurement.

## 7.5 nn-Meter: Complete Failure

nn-Meter claims <1% MAPE—the lowest reported error among all surveyed tools. After four deployment attempts (>4 hours), we obtained **zero predictions**. The failure mode is scikit-learn version incompatibility: pre-trained predictor models serialized with scikit-learn 0.23.1 (2020) cannot be deserialized with current versions. Available predictors cover Cortex-A76 CPU, Adreno 630/640 GPU, and Myriad VPU, but none are functional.

This result is significant: **the tool claiming the best accuracy is the only tool that produces no output**. nn-Meter’s failure demonstrates that pickle serialization of ML models without version pinning creates an expiration date on tool functionality—the tool became unusable within approximately two years of publication.

## 7.6 Cross-Cutting Findings

Our accuracy-centered evaluation reveals three findings that would be invisible to a survey based solely on self-reported numbers:

**First, self-reported accuracy is inversely correlated with tool reliability.** Ranking tools by claimed accuracy: nn-Meter (<1%) > NeuSight (2.3%) > VIDUR (<5%) > Timeloop (5–10%) > ASTRA-sim (5–15%). Ranking by actual reliability: VIDUR and ASTRA-sim (Docker-based, produce valid output in <30 minutes) > Timeloop (Docker, minor issues) > NeuSight (manual setup, accuracy overstated) > nn-Meter (completely broken). The tools claiming the lowest error are the least reliable in practice. This challenges the field’s implicit assumption that lower reported MAPE implies a better tool.

**Second, the five tools are complementary, not competing.** Table 4 reveals that no two tools meaningfully overlap on the same prediction task. NeuSight predicts GPU kernel latency; ASTRA-sim simulates multi-GPU communication; VIDUR models LLM serving; Timeloop explores accelerator design spaces; nn-Meter targets edge devices. The only partial overlap—NeuSight and Timeloop on single-layer latency—targets different hardware families (GPUs vs. systolic arrays). This disjointness means the field needs a *unified*

*pipeline* combining tool strengths, not a winner-take-all comparison (Section 8).

**Third, the composition gap dominates end-to-end prediction error.** NeuSight’s kernel-level accuracy is 5–9% MAPE on datacenter GPUs, but model-level prediction (summing kernel predictions) reaches 10–28% error on some configurations. The 5–15% additional error from composition—inter-kernel launch overhead, memory allocation, synchronization barriers—is *larger than most tools’ kernel-level error*. This means improving kernel predictors has diminishing returns until the composition problem is solved. No existing tool addresses this layer (Figure 4).

## 7.7 Threats to Validity

**External validity.** Our venue-focused search (ACM DL, IEEE Xplore, arXiv) may under-represent industry tools. We exclude proprietary tools (Nsight Compute [56], internal TPU models) from evaluation. The Apple M2 Ultra evaluation platform lacks discrete GPUs, preventing independent absolute accuracy verification for GPU-targeting tools.

**Internal validity.** Our evaluation covers 5 of 22 surveyed tools. Findings about specific methodology types rest on single tool instances—e.g., our assessment of ML-augmented approaches relies on nn-Meter, which may be unrepresentative due to its deployment failure. NeuSight’s accuracy analysis uses the tool’s own prediction/label pairs rather than independent hardware measurements; while this reveals consistency issues, hardware-verified measurements could differ.

**Construct validity.** Our accuracy-centered approach prioritizes a single dimension. Tools may provide value beyond accuracy—e.g., Timeloop’s energy breakdown is useful for design insight even if absolute numbers require hardware calibration. We partially address this through the feature availability matrix, but acknowledge that our evaluation is designed to challenge accuracy claims rather than comprehensively assess tool utility.

## 8 Toward a Unified Simulation Pipeline

The feature availability matrix (Table 4) reveals that the five evaluated tools cover fundamentally disjoint slices of the ML performance stack. No single tool predicts end-to-end performance from kernel execution through distributed training to serving-level SLAs. We propose a unified simulation pipeline that combines tool strengths across five layers, identifying the critical gaps that prevent integration today.

**Pipeline architecture.** The proposed pipeline composes predictions hierarchically:

- (1) **Hardware design** (Timeloop): For custom accelerators, explore the dataflow and mapping design space to determine per-layer energy and latency on a target architecture.
- (2) **Kernel prediction** (NeuSight / Timeloop): Predict per-kernel or per-layer execution time on the target GPU or accelerator. NeuSight covers 12 GPU types (NVIDIA + AMD); Timeloop covers systolic arrays and custom architectures.
- (3) **Model composition (CRITICAL GAP):** Compose kernel predictions into full model iteration time, accounting for inter-kernel launch overhead, memory allocation, data

- 813 movement between fused operator groups, and graph opti-  
 814 mization effects. No existing tool validates this layer.  
 815 (4) **Distributed training** (ASTRA-sim): Given per-device com-  
 816 pute time (from layers 1–3), simulate multi-GPU communica-  
 817 tion patterns, collective algorithms, and topology effects  
 818 to predict training throughput at scale.  
 819 (5) **Serving system** (VIDUR): For inference deployments, model  
 820 request-level scheduling, batching, KV cache management,  
 821 and queuing to predict TTFT, TPOT, and throughput under  
 822 realistic arrival patterns.

823 **Why combination is necessary.** ASTRA-sim models communica-  
 824 tion but not compute—it needs NeuSight for kernel times. VIDUR  
 825 models serving but uses profiled traces—it needs a predictor for un-  
 826 seen hardware. NeuSight predicts kernels but not system effects—it  
 827 needs ASTRA-sim for scale-out. Timeloop models accelerators but  
 828 not GPU workloads—it needs NeuSight for GPU targets. Each tool  
 829 fills a gap that the others cannot address.

830 **The critical gap: kernel-to-model composition.** The biggest  
 831 unsolved problem is Layer 3. NeuSight’s kernel-level accuracy is  
 832 5–9% MAPE on datacenter GPUs, but model-level error (naively  
 833 summing kernel predictions) reaches 10–28% on some configura-  
 834 tions. The composition gap of 5–15% additional error arises from  
 835 three sources: (1) kernel launch and scheduling overhead ( $\sim 5\text{--}10 \mu\text{s}$   
 836 per kernel), (2) inter-kernel data movement through the memory hi-  
 837 erarchy, and (3) synchronization barriers and stream management.  
 838 This gap is *larger than most tools’ kernel-level error*, meaning better  
 839 kernel predictors alone will not solve end-to-end accuracy.

840 **Integration requirements.** Realizing this pipeline requires:  
 841 (a) a common workload description format that all tools can con-  
 842 sume (currently each requires its own YAML/ONNX/trace format);  
 843 (b) validated composition models with formal error bounds; and  
 844 (c) cross-hardware accuracy transfer methods to maintain NeuSight-  
 845 level accuracy across GPU generations (currently, accuracy de-  
 846 grades 3–4× outside the training distribution).

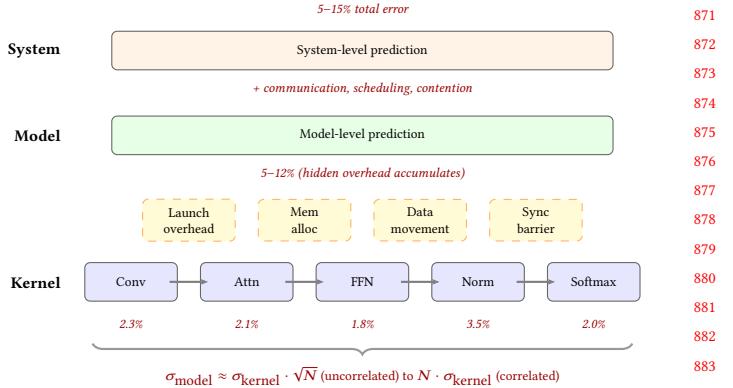
## 848 9 Open Challenges and Future Directions

850 Our accuracy-centered evaluation (Sections 6–8) exposes five con-  
 851 crete research directions, each grounded in empirical gaps.

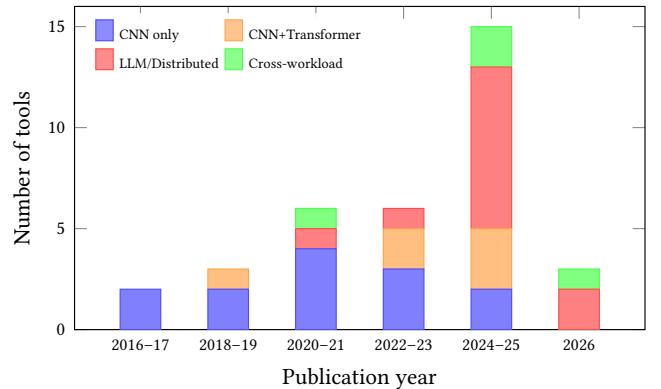
852 **1. Bridging the composition gap.** The composition problem  
 853 (Figure 4) is the field’s most pressing unsolved challenge. Kernel-  
 854 level errors of 2–3% yield ~5–12% model-level error through uncaptured  
 855 inter-kernel overheads ( $\sigma_{\text{model}} \approx \sigma_{\text{kernel}} \cdot \sqrt{N}$  for uncorrelated  
 856 errors, compounding linearly when correlated). No validated tool  
 857 pipeline exists from kernel prediction to system-level estimate; tools  
 858 either operate at a single level or sidestep composition by profiling  
 859 at the target level. Formal composition error bounds—analogous  
 860 to numerical error analysis—would enable practitioners to reason  
 861 about end-to-end accuracy from component specifications.

862 **2. Frontier workload coverage.** The temporal validation lag  
 863 (Section 4) is closing for transformers but remains wide: MoE, dif-  
 864 fusion [40], and dynamic inference lack validated tools; scaling  
 865 laws [14, 22, 27, 37] predict loss but not latency. Figure 5 shows the  
 866 post-2023 shift toward LLM workloads.

867 **3. Hardware transfer and emerging architectures.** Cross-  
 868 family transfer (GPU→TPU→PIM) remains unsolved despite meta-  
 869 learning (HELP) and feature-based transfer (LitePred). PIM [26, 31,



871 **Figure 4: Error composition across abstraction levels.** Kernel-  
 872 level predictions (2–3%) accumulate through unmodeled  
 873 inter-kernel overheads, yielding 5–12% model-level and 5–  
 874 15% system-level error.



891 **Figure 5: Workload coverage by publication period.** The shift  
 892 toward LLM workloads accelerates from 2023; MoE and dif-  
 893 fusion models remain uncharacterized.

906 46, 58], chiplets, and disaggregated designs blur memory hierarchy  
 907 assumptions that current analytical models rely on.

911 **4. Standardized evaluation infrastructure.** No MLPerf [53,  
 912 68] equivalent exists for performance *prediction*. Our accuracy-  
 913 centered approach provides a template; the community needs com-  
 914 mon benchmark suites, shared evaluation platforms, and standard-  
 915 ized reporting formats to make cross-tool comparison meaningful.  
 916 Portable workload formats (ONNX, Chakra [74]) and Docker-first  
 917 deployment are prerequisites.

918 **5. Temporal stability.** Software stack evolution (FlashAtten-  
 919 tion [16], new CUDA versions, framework updates) silently invali-  
 920 dates models. nn-Meter’s failure within two years demonstrates the  
 921 urgency; no tool currently addresses temporal robustness as a de-  
 922 sign goal. Future tools should adopt continuous validation against  
 923 evolving baselines [67].

## 929 10 Conclusion

930 This survey of 22 ML performance modeling tools provides accuracy-  
 931 centered independent evaluation of five tools through our own ex-  
 932 periments: 146 GPU configurations for NeuSight, collective bench-  
 933 marks for ASTRA-sim, LLM serving simulations for VIDUR, energy  
 934 validation for Timeloop, and reproducibility testing for nn-Meter.  
 935 Our evaluation yields three actionable findings. First, *self-reported*  
 936 *accuracy is unreliable*: NeuSight's claimed 2.3% MAPE is actually  
 937 5.87–27.10% depending on GPU, while nn-Meter (<1% claimed)  
 938 produces no output at all. Second, *the five tools are complemen-*  
 939 *tary rather than competing*—their feature coverage is fundamentally  
 940 disjoint—motivating a unified simulation pipeline that combines  
 941 kernel prediction, communication simulation, LLM serving, and  
 942 accelerator design into an end-to-end prediction framework. Third,  
 943 *the composition gap dominates end-to-end error*: the 5–15% gap be-  
 944 between kernel-level and model-level prediction is larger than kernel-  
 945 level error itself, meaning better kernel predictors have diminishing  
 946 returns until composition is solved. The most pressing community  
 947 needs are validated composition models, unified input formats  
 948 across tools, and continuous accuracy validation infrastructure.

## 951 References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 265–283.
- [2] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramachandran. 2024. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve. In *Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 117–134.
- [3] Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, and Ramachandran Ramachandran. 2024. VIDUR: A Large-Scale Simulation Framework for LLM Inference. In *Proceedings of Machine Learning and Systems (MLSys)*. 1–15.
- [4] Ali Bakhoda, George L. Yuan, Wilson W. L. Fung, Henry Wong, and Tor M. Aamodt. 2009. Analyzing CUDA Workloads Using a Detailed GPU Simulator. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 163–174. <https://doi.org/10.1109/ISPASS.2009.4919648>
- [5] Abhimanyu Rajeshkumar Bambhaniya et al. 2025. HERMES: Understanding and Optimizing Multi-Stage AI Inference Pipelines. *arXiv preprint arXiv:2504.09775* (2025). Heterogeneous multi-stage LLM inference simulator with analytical modeling.
- [6] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R. Hower, Tushar Krishna, Somayeh Sardashti, Rathijit Sen, Korey Sewell, Muhammad Shoaib, Nilay Vaish, Mark D. Hill, and David A. Wood. 2011. The gem5 Simulator. *ACM SIGARCH Computer Architecture News* 39, 2 (2011), 1–7. <https://doi.org/10.1145/2024716.2024718>
- [7] Shirley Browne, Jack Dongarra, Nathan Garner, George Ho, and Philip Mucci. 2000. A Portable Programming Interface for Performance Evaluation on Modern Processors. *International Journal of High Performance Computing Applications* 14, 3 (2000), 189–204. <https://doi.org/10.1177/10943420001400303> PAPI: portable API for hardware performance counters, foundational tool for performance analysis.
- [8] Kai Cai, Wei Miao, Junyu Zhu, Jiaxu Chen, Hao Shan, Huanyu Li, and Chi Zhang. 2024. Echo: Simulating Distributed Training At Scale. *arXiv preprint arXiv:2412.12487* (2024).
- [9] Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 1–15.
- [10] Zheng Cao et al. 2025. AMALI: An Analytical Model for Accurately Modeling LLM Inference on Modern GPUs. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA)*. 1–14. <https://doi.org/10.1145/3695053.3731064> Reduces GPU LLM inference MAPE from 127.56% to 23.59% vs GCoM baseline.
- [11] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning. In *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPOLOS)*. 269–284. <https://doi.org/10.1145/2541940.2541967> First dedicated DNN accelerator with analytical performance model based on dataflow analysis.
- [12] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Meghan Cowan, Haichen Shen, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. In *Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 578–594.
- [13] Yu-Hsin Chen, Joel Emer, and Vivienne Sze. 2016. Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks. In *Proceedings of the 43rd International Symposium on Computer Architecture (ISCA)*. 367–379. <https://doi.org/10.1109/ISCA.2016.40>
- [14] Leshem Choshen, Yang Zhang, and Jacob Andreas. 2025. A Hitchhiker's Guide to Scaling Law Estimation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. 1–25. Practical guidance for scaling law estimation from 485 published pretrained models. IBM/MIT.
- [15] Weiwei Chu, Xinfeng Xie, Jiecao Yu, Jie Wang, Pavan Balaji, Ching-Hsiang Chu, Jongsoo Park, et al. 2025. Scaling Llama 3 Training with Efficient Parallelism Strategies. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA)*. 1–15. 4D parallelism for Llama 3 405B on 16K H100 GPUs. Achieves 400 TFLOPS/GPU. Meta.
- [16] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 16344–16359.
- [17] Lukasz Dudziak, Thomas Chau, Mohamed S. Abdelfattah, Royson Lee, Hyeji Kim, and Nicholas D. Lane. 2024. Latency Predictors for Neural Architecture Search. In *Proceedings of Machine Learning and Systems (MLSys)*. 1–14.
- [18] Yang Feng, Zhehae Li, Jiacheng Yang, and Yunxin Liu. 2024. LitePred: Transferable and Scalable Latency Prediction for Hardware-Aware Neural Architecture Search. In *Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. 1–18.
- [19] Paraskevas Gavriilidis et al. 2025. LIFE: Forecasting LLM Inference Performance via Hardware-Agnostic Analytical Modeling. *arXiv preprint arXiv:2508.00904* (2025). Hardware-agnostic analytical model for LLM inference performance forecasting.
- [20] Siddharth Ghosh et al. 2025. Frontier: Simulating the Next Generation of LLM Inference Systems. *arXiv preprint arXiv:2508.03148* (2025). Stage-centric simulator for MoE and disaggregated LLM inference, models expert parallelism and cross-cluster routing.
- [21] Alicia Golden et al. 2025. PRISM: Probabilistic Runtime Insights and Scalable Performance Modeling for Large-Scale Distributed Training. *arXiv preprint arXiv:2510.15596* (2025). Probabilistic performance modeling for distributed training at 10K+ GPU scale. Meta.
- [22] Alexander Hagelé, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. 2024. Scaling Laws and Compute-Optimal Training Beyond Fixed Training Durations. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37. Spotlight. Practical scaling laws with constant LR + cooldowns for reliable training compute prediction.
- [23] Amer Haj-Ali et al. 2025. Omnidrive: Predicting GPU Kernels Performance with LLMs. *arXiv preprint arXiv:2506.20886* (2025). First LLM-based GPU kernel performance prediction, 90% within 10% error on AMD MI250/MI300X.
- [24] Yanbin Hao et al. 2025. POD-Attention: Unlocking Full Prefill-Decode Overlap for Faster LLM Inference. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (AS- PLOS)*. 1–15. Full overlap between prefill and decode phases for LLM inference.
- [25] John L. Hennessy and David A. Patterson. 2019. A New Golden Age for Computer Architecture. *Commun. ACM* 62, 2 (2019), 48–60. <https://doi.org/10.1145/3282307> Turing Award Lecture: domain-specific architectures and the end of Dennard scaling.
- [26] Guseul Heo, Sangyeop Lee, Jaehong Cho, Hyunmin Choi, Sanghyeon Lee, Hyungkyu Ham, Gwangsun Kim, Divya Mahajan, and Jongse Park. 2024. Ne-upIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inference. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (AS- PLOS)*. 1–17. NPU-PIM heterogeneous architecture for LLM inference with performance modeling. KAIST/Georgia Tech.
- [27] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556* (2022). Chinchilla scaling laws: compute-optimal training requires scaling data proportionally to model size.
- [28] Samuel Hsia, Kartik Chandra, and Kunle Olukotun. 2024. MAD Max Beyond Single-Node: Enabling Large Machine Learning Model Acceleration on Distributed Systems. In *Proceedings of the 51st Annual International Symposium on*

- 1045        Computer Architecture (ISCA). 753–766. <https://doi.org/10.1109/ISCA59077.2024.00064>
- 1046 [29] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Xu  
1047 Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng  
1048 Chen. 2019. GPipe: Efficient Training of Giant Neural Networks using Pipeline  
1049 Parallelism. In *Advances in Neural Information Processing Systems (NeurIPS)*,  
1050 Vol. 32. 103–112.
- 1051 [30] Rodrigo Huerta, Mojtaba Abae Shoushtary, Jose-Lorenzo Cruz, and Antonio  
1052 Gonzalez. 2025. Dissecting and Modeling the Architecture of Modern GPU Cores.  
1053 In *Proceedings of the 58th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 369–384. Reverse-engineers modern NVIDIA GPU cores, improves  
1054 Accel-Sim to 13.98% MAPE. UPC Barcelona.
- 1055 [31] Bongjoon Hyun, Taejun Kim, Dongjae Lee, and Minsoo Rhu. 2024. Pathfinding Future PIM Architectures by Demystifying a Commercial PIM Technology.  
1056 In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 1–15. uPIMulator: cycle-accurate PIM simulation  
1057 framework for UPMEM. KAIST.
- 1058 [32] Ryota Imai, Kentaro Harada, Ryo Sato, and Toshio Nakaike. 2024. Roofline-  
1059 Driven Machine Learning for Large Language Model Performance Prediction.  
1060 *NeurIPS Workshop on Machine Learning for Systems* (2024).
- 1061 [33] Anand Jayaraman, Wei-Lin Hu, Gauri Zhao, and Gennady Pekhimenko. 2023.  
1062 Sia: Heterogeneity-aware, Goodput-optimized ML-Cluster Scheduling. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. 642–657.  
1063 <https://doi.org/10.1145/3600006.3613175> Extends goodput optimization to het-  
1064 erogeneous GPU clusters for training workloads.
- 1065 [34] Norman P. Jouppi, Doe Hyun Yoon, George Kurian, Sheng Li, Nishant Patil,  
1066 James Laudon, Cliff Young, and David Patterson. 2023. TPU v4: An Optically  
1067 Reconfigurable Supercomputer for Machine Learning with Hardware Support for  
1068 Embeddings. *Proceedings of the 50th Annual International Symposium on Computer  
1069 Architecture (ISCA)* (2023), 1–14. <https://doi.org/10.1145/3579371.3589350> 4096-  
1070 chip pods with 3D optical interconnect; up to 1.7x/2.1x faster than TPU v3.
- 1071 [35] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal,  
1072 Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borber, et al. 2017.  
1073 In-Datcenter Performance Analysis of a Tensor Processing Unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*.  
1074 1–12. <https://doi.org/10.1145/3079856.3080246> First dedicated ML inference  
1075 accelerator; 15–30x over CPUs/GPUs on CNN inference.
- 1076 [36] Andreas Kosmas Kakolyris, Dimosthenis Masouros, Petros Vavaroutsos, Sotirios  
1077 Xydis, and Dimitrios Soudris. 2025. throtLL'eM: Predictive GPU Throttling for  
1078 Energy Efficient LLM Inference Serving. In *Proceedings of the IEEE International  
1079 Symposium on High Performance Computer Architecture (HPCA)*. 1–14. Achieves  
1080 up to 43.8% lower energy consumption for LLM inference.
- 1081 [37] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess,  
1082 Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020.  
1083 Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*  
1084 (2020). Original neural scaling laws: power-law relationships between model  
1085 size, dataset size, compute, and loss.
- 1086 [38] Mahmoud Khairy, Zhesheng Shen, Tor M. Aamodt, and Timothy G. Rogers. 2020.  
1087 Accel-Sim: An Extensible Simulation Framework for Validated GPU Modeling. In  
1088 *Proceedings of the 47th International Symposium on Computer Architecture (ISCA)*.  
1089 473–486. <https://doi.org/10.1109/ISCA45697.2020.00047>
- 1090 [39] Jungho Kim et al. 2025. PyTorchSim: A Comprehensive, Fast, and Accurate  
1091 NPU Simulation Framework. In *Proceedings of the 58th IEEE/ACM International  
1092 Symposium on Microarchitecture (MICRO)*. 1–14. <https://doi.org/10.1145/3725843.3756045> PyTorch 2-integrated NPU simulator with custom RISC-V ISA and  
1093 Tile-Level Simulation.
- 1094 [40] Jin Kim, Byeongjun Shin, Jinha Chung, and Minsoo Rhu. 2026. The Cost of  
1095 Dynamic Reasoning: Demystifying AI Agents and Test-Time Scaling from an AI  
1096 Infrastructure Perspective. In *Proceedings of the IEEE International Symposium  
1097 on High Performance Computer Architecture (HPCA)*. 1–14. HPCA 2026 (Jan 31–  
1098 Feb 4, 2026, Las Vegas). First comprehensive system-level analysis of AI agents;  
1099 quantifies resource usage, latency, and datacenter power consumption.
- 1100 [41] Yoongi Kim, Weikun Yang, and Onur Mutlu. 2016. Ramulator: A Fast and  
1101 Extensible DRAM Simulator. *IEEE Computer Architecture Letters* 15, 1 (2016), 45–  
1102 49. <https://doi.org/10.1109/LCA.2015.2414456> Fast extensible DRAM simulator  
1103 supporting DDRx, LPDDRx, GDDRx, WIOx, HBMx standards.
- 1104 [42] Srivatsan Krishnan, Amir Yazdanbakhsh, Shvetank Prakash, Norman P.  
1105 Jouppi, Jignesh Parmar, Hyoukjun Kim, James Laudon, and Chandrakant  
1106 Narayanaswami. 2023. ArchGym: An Open-Source Gymnasium for Machine  
1107 Learning Assisted Architecture Design. In *Proceedings of the 50th International  
1108 Symposium on Computer Architecture (ISCA)*. 1–16. <https://doi.org/10.1145/3579371.3589049>
- 1109 [43] Hyoukjun Kwon, Prasanth Chatarasi, Michael Sarber, Michael Pellauer, Angshuman  
1110 Parashar, and Tushar Krishna. 2019. MAESTRO: A Data-Centric Approach  
1111 to Understand Reuse, Performance, and Hardware Cost of DNN Mappings. In  
1112 *Proceedings of the 52nd IEEE/ACM International Symposium on Microarchitecture  
1113 (MICRO)*. 1–14. <https://doi.org/10.1145/3352460.3358292>
- 1114 [44] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng,  
1115 Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient  
1116 Memory Management for Large Language Model Serving with PagedAttention.  
1117 In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*.  
1118 611–626. <https://doi.org/10.1145/3600006.3613165>
- 1119 [45] Chris Lattner, Mehdi Amini, Uday Bondhugula, Albert Cohen, Andy Davis,  
1120 Jacques Pienaar, River Riddle, Tatiana Shpeisman, Nicolas Vasilache, and Oleksandr  
1121 Zinenko. 2021. MLIR: Scaling Compiler Infrastructure for Domain Specific  
1122 Computation. In *Proceedings of the IEEE/ACM International Symposium on Code  
1123 Generation and Optimization (CGO)*. 2–14. <https://doi.org/10.1109/CGO51591.2021.9370308> Multi-level IR infrastructure enabling cost model composition  
1124 across abstraction levels.
- 1125 [46] Hyojung Lee, Daehyeon Baek, Jimyoung Son, Jieun Choi, Kihyo Moon, and  
1126 Minsung Jang. 2025. PAISE: PIM-Accelerated Inference Scheduling Engine for  
1127 Transformer-based LLM. In *Proceedings of the IEEE International Symposium  
1128 on High Performance Computer Architecture (HPCA)*. 1–14. PIM-based LLM  
1129 inference scheduling, 48.3% speedup, 11.5% power reduction. Samsung.
- 1130 [47] Hayeon Lee, Sewoong Lee, Song Chong, and Sung Ju Hwang. 2021. HELP:  
1131 Hardware-Adaptive Efficient Latency Prediction for NAS via Meta-Learning.  
1132 In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. 27016–  
1133 27028.
- 1134 [48] Seunghyun Lee, Amar Phanishayee, and Divya Mahajan. 2025. NeuSight: GPU  
1135 Performance Forecasting via Tile-Based Execution Analysis. In *Proceedings of the  
1136 30th ACM International Conference on Architectural Support for Programming  
1137 Languages and Operating Systems (ASPLOS)*. 1–15.
- 1138 [49] Jianbo Li et al. 2025. TrioSim: A Lightweight Simulator for Large-Scale DNN  
1139 Workloads on Multi-GPU Systems. In *Proceedings of the 52nd Annual International  
1140 Symposium on Computer Architecture (ISCA)*. 1–13. Multi-GPU DNN simulation  
1141 with lightweight approach for distributed training analysis.
- 1142 [50] Shang Li, Zhiyuan Yang, Dhiraj Reddy, Ankur Srivastava, and Bruce Jacob. 2020.  
1143 DRAMsim3: A Cycle-Accurate, Thermal-Capable DRAM Simulator. *IEEE Computer  
1144 Architecture Letters* 19, 2 (2020), 106–109. <https://doi.org/10.1109/LCA.2020.2973991> Modernized DRAM simulator with thermal modeling and HMC  
1145 support.
- 1146 [51] Wenxuan Liang et al. 2025. Lumos: Efficient Performance Modeling and Esti-  
1147 mation for Large-scale LLM Training. In *Proceedings of Machine Learning and  
1148 Systems (MLSys)*. 1–16. Trace-driven performance modeling achieving 3.3% error  
1149 on H100 GPUs for LLM training.
- 1150 [52] Haocong Luo, Yahya Can Tugrul, F. Nisa Bostancı, Ataberk Olgun, A. Giray  
1151 Yaghkci, and Onur Mutlu. 2023. Ramulator 2.0: A Modern, Modular, and Extensi-  
1152 ble DRAM Simulator. *IEEE Computer Architecture Letters* 22, 2 (2023), 129–132.  
1153 <https://doi.org/10.1109/LCA.2023.3333759> Modular DRAM simulator with DDR5,  
1154 LPDDR5, HBM3, GDDR6 support and RowHammer mitigation modeling.
- 1155 [53] Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micke-  
1156 vicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf,  
1157 et al. 2020. MLPerf Training Benchmark. In *Proceedings of Machine Learning  
1158 and Systems (MLSys)*. 336–349. Standard ML training benchmark suite cover-  
1159 ing image classification, object detection, NLP, recommendation, reinforcement  
1160 learning.
- 1161 [54] Azaz-Ur-Rehman Nasir, Samroz Ahmad Shoib, Muhammad Abdullah Hanif, and  
1162 Muhammad Shafique. 2025. ESM: A Framework for Building Effective Surrogate  
1163 Models for Hardware-Aware Neural Architecture Search. In *Proceedings of the  
1164 62nd ACM/IEEE Design Automation Conference (DAC)*. 1–6. 97.6% accuracy  
1165 surrogate model framework for HW-aware NAS.
- 1166 [55] Amir Nasr-Esfahany et al. 2025. Concorde: Fast and Accurate CPU Performance  
1167 Modeling with Compositional Analytical-ML Fusion. In *Proceedings of the 52nd  
1168 Annual International Symposium on Computer Architecture (ISCA)*. 1–15. Hybrid  
1169 analytical-ML approach achieving 2% CPI error at 5 orders of magnitude faster  
1170 than gem5.
- 1171 [56] NVIDIA Corporation. 2019. Nsight Compute: Interactive Kernel Profiler. <https://developer.nvidia.com/nsight-compute>. Industry-standard GPU kernel profiling  
1172 tool with roofline analysis.
- 1173 [57] Angshuman Parashar, Priyanka Raina, Yakun Sophia Shao, Yu-Hsin Chen,  
1174 Victor A. Ying, Anurag Muber, Rangharajan Venkatesan, Brueck Khalilany,  
1175 Stephen W. Keckler, and Joel Emer. 2019. Timeloop: A Systematic Approach  
1176 to DNN Accelerator Evaluation. In *Proceedings of the IEEE International Sym-  
1177 posium on Performance Analysis of Systems and Software (ISPASS)*. 304–315.  
1178 <https://doi.org/10.1109/ISPASS.2019.00042>
- 1179 [58] Jaehyun Park, Jaewan Choi, Kwanhee Kyung, Michael Jaemin Kim, Yongsuk  
1180 Kwon, Nam Sung Kim, and Jung Ho Ahn. 2024. AttAcc! Unleashing the Power of  
1181 PIM for Batched Transformer-based Generative Model Inference. In *Proceedings of  
1182 the 29th ACM International Conference on Architectural Support for Programming  
1183 Languages and Operating Systems (ASPLOS)*. 1–16. PIM-based accelerator for  
1184 batched transformer attention. Seoul National University/UIUC..
- 1185 [59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory  
1186 Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019.  
1187 PyTorch: An Imperative Style, High-Performance Deep Learning Library. In  
1188 *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 32. 8024–8035.

- [60] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aakanksha Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient Generative LLM Inference Using Phase Splitting. In *Proceedings of the 51st Annual International Symposium on Computer Architecture (ISCA)*. 118–132. <https://doi.org/10.1109/ISCA59077.2024.00019> Best Paper Award.
- [61] Hang Qi, Evan R. Sparks, and Ameet Talwalkar. 2017. Paleo: A Performance Model for Deep Neural Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=SyVvJ85lg>
- [62] Aurick Qiao, Sang Keun Agrawal, Anand Jayaraman, Moustafa Mittal, Amar Altaf, Michael Cho, and Gennady Pekhimenko. 2021. Pollux: Co-adaptive Cluster Scheduling for Goodput-Optimized Deep Learning. In *Proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 1–18. Goodput estimation for co-optimizing resource allocation and training hyperparameters.
- [63] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédéric Durand, and Saman Amarasinghe. 2013. Halide: A Language and Compiler for Optimizing Parallelism, Locality, and Recomputation in Image Processing Pipelines. In *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*. 519–530. <https://doi.org/10.1145/2491956.2462176> Pioneered separation of algorithm and schedule with learned cost models for autoscheduling.
- [64] Samyam Rajbhandari, Jeff Rasley, Olatunji Rber, and Yuxiong He. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC)*. 1–16. <https://doi.org/10.1109/SC41405.2020.00024> DeepSpeed ZeRO optimizer partitioning for memory-efficient distributed training.
- [65] Mehdi Rakhshanfar and Aliakbar Zarandi. 2021. A Survey on Machine Learning-based Design Space Exploration for Processor Architectures. *Journal of Systems Architecture* 121 (2021), 102339. <https://doi.org/10.1016/j.sysarc.2021.102339>
- [66] Saeed Rashidi, Srinivas Srinivasan, Kazem Hamedani, and Tushar Krishna. 2020. ASTRA-SIM: Enabling SW/HW Co-Design Exploration for Distributed DL Training Platforms. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 81–92. <https://doi.org/10.1109/ISPASS48437.2020.00018>
- [67] Vijay Janapa Reddi et al. 2025. MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Inference. In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 1–14. Energy efficiency benchmarking for ML inference workloads.
- [68] Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, Brian Anderson, Maxim Breeshevov, Mark Duber, et al. 2020. MLPerf Inference Benchmark. In *Proceedings of the 47th International Symposium on Computer Architecture (ISCA)*. 446–459. <https://doi.org/10.1109/ISCA45697.2020.00045> Standard ML inference benchmark suite with server and offline scenarios.
- [69] Arun F. Rodrigues, K. Scott Hemmert, Brian W. Barrett, Chad Kersey, Ron Oldfield, Marlo Weston, R. Risen, Jeanine Cook, Paul Rosenfeld, Elliott Cooper-Balis, and Bruce Jacob. 2012. The Structural Simulation Toolkit. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 38. 37–42. <https://doi.org/10.1145/1964218.1964225> Modular framework for system-level simulation, widely used for HPC and interconnect modeling.
- [70] Paul Rosenfeld, Elliott Cooper-Balis, and Bruce Jacob. 2011. DRAMSim2: A Cycle Accurate Memory System Simulator. *IEEE Computer Architecture Letters* 10, 1 (2011), 16–19. <https://doi.org/10.1109/L-CA.2011.4> Widely-used cycle-accurate DDR/DDR3 memory simulator validated against manufacturer Verilog models.
- [71] Ananda Samajdar, Yuhao Zhu, Paul Whatmough, Matthew Mattina, and Tushar Krishna. 2019. A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-Sim. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 58–68. <https://doi.org/10.1109/ISPASS.2019.00016> Cycle-accurate systolic array simulator for DNN accelerator DSE.
- [72] Zhuomin Shen, Jaeho Kim, et al. 2025. AQUA: Network-Accelerated Memory Offloading for LLMs in Scale-Up GPU Domains. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPiOS)*. 1–16. <https://doi.org/10.1145/3676641.3715983> Improves LLM inference responsiveness by 20x through network-accelerated memory offloading.
- [73] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. In *arXiv preprint arXiv:1909.08053*. Intra-layer tensor parallelism for large language model training.
- [74] Srinivas Sridharan, Taekyung Heo, Jinwoo Choi, Garyfallia Yu, Saeed Rashidi, William Won, Zhaodong Meng, and Tushar Krishna. 2023. Chakra: Advancing Performance Benchmarking and Co-design using Standardized Execution Traces. *arXiv preprint arXiv:2305.14516* (2023).
- [75] Foteini Strati, Zhendong Zhang, George Manos, Ixeia Sanchez Periz, Qinghao Hu, Tiancheng Chen, Berk Buzcu, Song Han, Pamela Delgado, and Ana Klimovic. 2025. Sailor: Automating Distributed Training over Dynamic, Heterogeneous, and Geo-distributed Clusters. In *Proceedings of the 30th ACM Symposium on Operating Systems Principles (SOSP)*. 1–18. Automated distributed training with runtime/memory simulation over heterogeneous resources. ETH Zurich/MIT.
- [76] Ondrej Sýkora, Alexis Rucker, Charith Mendis, Rajkishore Barik, Phitchaya Phothilimthana, and Saman Amarasinghe. 2022. GRANITE: A Graph Neural Network Model for Basic Block Throughput Estimation. In *Proceedings of the IEEE International Symposium on Workload Characterization (IISWC)*. 1–13. <https://doi.org/10.1109/IISWC55918.2022.00014>
- [77] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. 2017. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. In *Proceedings of the IEEE*, Vol. 105. 2295–2329. <https://doi.org/10.1109/JPROC.2017.2761740> Canonical DNN accelerator taxonomy covering dataflows, data reuse, and energy efficiency.
- [78] Philippe Tillet, H. T. Kung, and David Cox. 2019. Triton: An Intermediate Language and Compiler for Tiled Neural Network Computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages (MAPL)*. 10–19. <https://doi.org/10.1145/3315508.3329973> Tile-based GPU programming with heuristic performance model for kernel generation.
- [79] Jan Treibig, Georg Hager, and Gerhard Wellein. 2010. LIKWID: A Lightweight Performance-Oriented Tool Suite for x86 Multicore Environments. In *Proceedings of the 39th International Conference on Parallel Processing Workshops (ICPPW)*. 207–216. <https://doi.org/10.1109/ICPPW.2010.38> Lightweight tools for thread/cache topology, affinity, and performance counter measurement.
- [80] Adrian Tschandl, Mohamed Awad, et al. 2025. SwizzlePerf: Hardware-Aware LLMs for GPU Kernel Performance Optimization. *arXiv preprint arXiv:2508.20258* (2025). LLM-based spatial optimization for GPU kernels, up to 2.06x speedup via swizzling.
- [81] Xizheng Wang, Qingxu Li, Yichi Xu, Gang Lu, Heyang Zhou, Sen Zhang, Yikai Zhu, Yang Liu, Pengcheng Zhang, Kun Qian, et al. 2025. SimAI: Unifying Architecture Design and Performance Tuning for Large-Scale LLM Training with Scalability and Precision. In *Proceedings of the 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. 1–18. Full-stack LLM training simulator achieving 98.1% alignment with real-world results. Alibaba Cloud/Tsinghua.
- [82] Zixian Wang et al. 2025. SynPerf: Synthesizing High-Performance GPU Kernels via Pipeline Decomposition. *arXiv preprint* (2025). Under review.
- [83] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: An Insightful Visual Performance Model for Multicore Architectures. *Commun. ACM* 52, 4 (2009), 65–76. <https://doi.org/10.1145/1498765.1498785>
- [84] William Won, Taekyung Heo, Saeed Rashidi, Saeed Talati, Srinivas Srinivasan, and Tushar Krishna. 2023. ASTRA-sim2.0: Modeling Hierarchical Networks and Disaggregated Systems for Large-Model Training at Scale. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 283–294. <https://doi.org/10.1109/ISPASS57527.2023.00035>
- [85] Yannan Nellie Wu, Joel Emer, and Vivienne Sze. 2022. Sparseloop: An Analytical Approach to Sparse Tensor Accelerator Modeling. In *Proceedings of the 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 1–15. <https://doi.org/10.1109/MICRO56248.2022.00078>
- [86] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. ORCA: A Distributed Serving System for Transformer-Based Generative Models. In *Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 521–538.
- [87] Geoffrey X. Yu, Yubo Gao, Pavel Golber, and Asaf Cidon. 2021. Habitat: A Runtime-Based Computational Performance Predictor for Deep Neural Network Training. In *Proceedings of the USENIX Annual Technical Conference (ATC)*. 503–521.
- [88] Yi Zhai, Yu Cheng Wang, Peng Jiang, and Congming Kang. 2023. TLP: A Deep Learning-based Cost Model for Tensor Program Tuning. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. 833–845. <https://doi.org/10.1145/3575693.3575736>
- [89] Li Lyra Zhang, Shihao Han, Jianyu Wei, Ningxin Zheng, Ting Cao, Yuqing Yang, and Yunxin Liu. 2021. nn-Meter: Towards Accurate Latency Prediction of Deep-Learning Model Inference on Diverse Edge Devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 81–93. <https://doi.org/10.1145/3458864.3467882> Best Paper Award.
- [90] Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, Amer Haj-Ali, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, Joseph E. Gonzalez, and Ion Stoica. 2020. Ansor: Generating High-Performance Tensor Programs for Deep Learning. In *Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 863–879.
- [91] Lianmin Zheng, Ruochen Liu, Junru Shao, Tianqi Chen, Joseph E. Gonzalez, Ion Stoica, and Zhihao Zhang. 2021. TenSet: A Large-scale Program Performance Dataset for Learned Tensor Compilers. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. 29876–29888.

1277	[92] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianyu Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. 2024. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving. In <i>Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)</i> . 1–18.	1335
1278		1336
1279		1337
1280		1338
1281		1339
1282		1340
1283		1341
1284		1342
1285		1343
1286		1344
1287		1345
1288		1346
1289		1347
1290		1348
1291		1349
1292		1350
1293		1351
1294		1352
1295		1353
1296		1354
1297		1355
1298		1356
1299		1357
1300		1358
1301		1359
1302		1360
1303		1361
1304		1362
1305		1363
1306		1364
1307		1365
1308		1366
1309		1367
1310		1368
1311		1369
1312		1370
1313		1371
1314		1372
1315		1373
1316		1374
1317		1375
1318		1376
1319		1377
1320		1378
1321		1379
1322		1380
1323		1381
1324		1382
1325		1383
1326		1384
1327		1385
1328		1386
1329		1387
1330		1388
1331		1389
1332		1390
1333		1391
1334		1392