

DSD: A DISTRIBUTED SPECULATIVE DECODING SOLUTION FOR EDGE-CLOUD AGILE LARGE MODEL SERVING

Fengze Yu¹ Leshu Li² Brad McDanel³ Saiqian Zhang¹

ABSTRACT

Large language model (LLM) inference often suffers from high decoding latency and limited scalability across heterogeneous edge–cloud environments. Existing speculative decoding (SD) techniques accelerate token generation but remain confined to single-node execution. We propose *DSD*, a distributed speculative decoding framework that extends SD to multi-device deployments through coordinated draft–target execution. Given the lack of prior work on simulating this paradigm, we first introduce *DSD-Sim*, a discrete-event simulator that captures network, batching, and scheduling dynamics. Building on insights from DSD-Sim, we further design an *Adaptive Window Control* (AWC) policy that dynamically adjusts speculation window size to optimize throughput. Experiments across diverse workloads show that DSD achieves up to **1.1**× speedup and **9.7%** higher throughput over existing SD baselines, enabling agile and scalable LLM serving across edge and cloud.

1 INTRODUCTION

Large Language Models (LLMs) have become a cornerstone of modern artificial intelligence by advancing natural language understanding and generation. These models enable machines to process, interpret, and produce human-like text, supporting a wide range of applications across domains. Despite their impressive capabilities, LLMs continue to suffer from significant processing latency, primarily due to the computationally intensive prefilling and autoregressive decoding stages. To address this, speculative decoding (SD) (Stern et al., 2018; Leviathan et al., 2023) has emerged as an effective acceleration strategy. SD decomposes the autoregressive generation process into two stages: a lightweight drafting stage that proposes multiple candidate tokens, followed by a verification stage in which the target LLM evaluates these candidates in parallel. This two-phase design improves decoding throughput while preserving output quality via a selective acceptance–rejection mechanism.

However, although SD can substantially accelerate LLM inference, it remains insufficient to meet the scalability requirements of large-scale or real-time deployments. SD primarily improves token generation efficiency within a single device or limited hardware environment. Yet, modern LLM applications—such as interactive AI chatbots (OpenAI, 2023), intelligent tutoring systems (Kasneci et al., 2023),

¹New York University, New York, NY, USA ²University of Minnesota Twin Cities, Minneapolis, MN, USA ³Franklin & Marshall College, Lancaster, PA, USA. Correspondence to: Saiqian Zhang <sai.zhang@nyu.edu>.

and AI recommendation systems (Gao et al., 2023)—often need to serve thousands of concurrent users while maintaining low latency.

Crucially, most existing SD frameworks focus on algorithmic efficiency and assume that the draft and target models reside on the same device (Chen et al., 2023) or, at most, across two nodes (Liu et al., 2024). As a result, they remain limited by the inherent sequential dependency between the drafting and verification stages, which constrains overall throughput. Moreover, the large computational footprint of the target LLM makes it impractical to deploy within resource-constrained edge environments, further restricting SD’s applicability in real-world edge–cloud scenarios.

As model complexity and application demands escalate, achieving true scalability necessitates extending SD beyond traditional single-node or small-scale deployments. A distributed SD framework can leverage both model and data parallelism to offload and coordinate computation across edge and cloud resources, thereby reducing latency, improving resource utilization, and enabling efficient large-scale LLM inference for real-world applications.

An example deployment scenario is illustrated in Figure 1 (a). Given the large scale of the target model, it is hosted across M cloud servers responsible for performing the verification stage on behalf of N edge LLMs, each serving as a draft model on resource-constrained edge devices. During operation, an edge LLM receives a user request and context prompt, generates a set of γ draft tokens, and transmits them to the cloud LLM for verification. Based on the verification feedback, the edge LLM continues generating subsequent

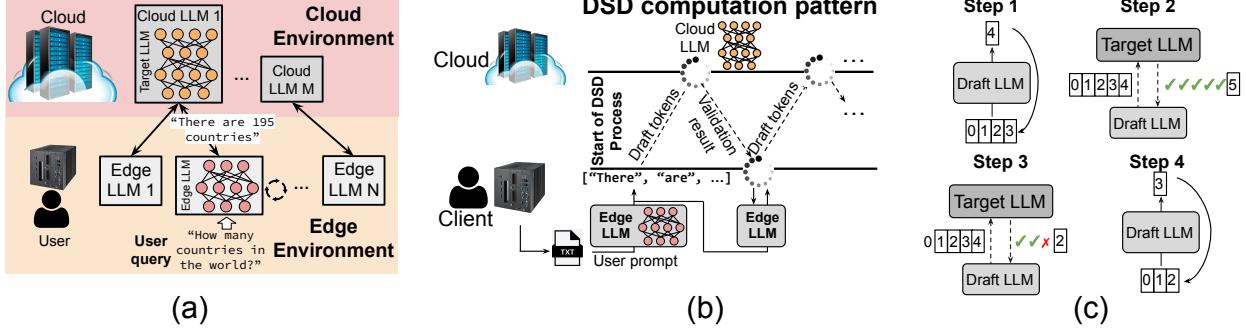


Figure 1. (a) Distributed edge–cloud environment for speculative decoding. (b) Joint processing between edge and cloud servers during SD operation. (c) Illustration of the speculative decoding workflow.

tokens until the completion of the response sequence. The process is described in Figure 1 (b).

In this paper, we introduce *DSD*, a unified distributed speculative decoding framework designed to enable scalable LLM serving across multiple nodes. Our contributions can be summarized as follows:

- We develop *DSD-Sim*, a large-scale DSD simulation framework that accurately models the performance behavior of distributed speculative decoding systems under diverse system and network conditions. *DSD-Sim* follows a modular design philosophy and is compatible with a wide range of existing simulation frameworks for LLM inference, networking, and distributed computing, allowing seamless integration and extensibility.
- Motivated by the performance bottlenecks identified through *DSD-Sim*, we further introduce an *Adaptive Window Control* (AWC) mechanism to enhance the overall throughput of the DSD framework. AWC employs a deep learning–based approach that dynamically adjusts the window configuration based on real-time system and algorithmic conditions, enabling the system to maintain optimal performance across varying workloads and network environments.
- Extensive experiments on GSM8K, CNN/DailyMail, and HumanEval benchmarks show that DSD combined with AWC consistently enhances system efficiency, achieving up to 9.7% higher throughput and 11% lower TPOT than fixed-threshold and heuristic scheduling baselines, all without manual tuning.

2 BACKGROUND AND RELATED WORK

2.1 Speculative Decoding

Speculative decoding is an effective technique for mitigating the sequential bottleneck in language model inference (Stern et al., 2018; Chen et al., 2023). It decomposes the decoding

process into two stages: a lightweight draft model quickly proposes a sequence of candidate tokens, which are then verified in parallel by a more accurate target model, as illustrated in Figure 1 (c).

Let the draft model M_{da} generate γ draft tokens (t_1, \dots, t_γ) in each iteration (step 1 of Figure 1 (c)), where γ denotes the *speculation window size*. During verification, the target model M_{ta} evaluates all draft tokens in parallel but accepts them sequentially. If all tokens are accepted, M_{da} proceeds to produce the next batch of candidates (step 2 of Figure 1 (c)). Otherwise, if a mismatch occurs at position i , all tokens from t_i onward are discarded, and the target model’s own sampled token t'_i is used instead (step 3 of Figure 1 (c)). The resulting accepted sequence is therefore $(t_1, \dots, t_{i-1}, t'_i)$, and the draft model will generate subsequent tokens based on the prefix determined by M_{ta} (step 4 of Figure 1 (c)).

Assuming a per-token acceptance probability α , the expected number of accepted draft tokens is:

$$\mathbb{E}[\tau] = \frac{1 - \alpha^{\gamma+1}}{1 - \alpha}. \quad (1)$$

Let c denote the cost ratio between the draft and target models per token. The expected speedup over standard decoding with M_{ta} is then:

$$S = \frac{1 - \alpha^{\gamma+1}}{(1 - \alpha)(c\gamma + 1)}. \quad (2)$$

This acceptance mechanism is compatible with various sampling temperatures. By allowing parallel draft generation while maintaining output fidelity through sequential verification, speculative decoding effectively relaxes the step-by-step dependency inherent in traditional autoregressive decoding.

2.2 Split Computing

Split neural network computing has attracted substantial attention from both academia and industry, as reflected in

a broad range of studies (Hauswald et al., 2014; Teerapittayanan et al., 2016; Kang et al., 2017; Teerapittayanan et al., 2017; Karjee et al., 2022; Luo et al., 2023; Mubark et al., 2024; Lee et al., 2023; Feltin et al., 2023; Zeng et al., 2020; Ding et al., 2023; Kang et al., 2022; Matsubara et al., 2019; 2022; Lin et al., 2024; Zhang et al., 2022; Dong et al., 2022; Zhang et al., 2024). Both commercial and open-source systems have adopted split learning and distributed inference in practice (pys, 2020; spl, 2020).

Among various partitioning strategies (Kang et al., 2017; Zhang et al., 2020), layer-wise partitioning is the most widely adopted (Hauswald et al., 2014; Teerapittayanan et al., 2016; Kang et al., 2017; Teerapittayanan et al., 2017). In this approach, a deep neural network (DNN) is divided into two or more segments that are executed collaboratively across multiple devices.

One of the earliest works in this direction, Hauswald *et al.* (Hauswald et al., 2014), proposed offloading the later stages of image classification computation to cloud servers. Following this, Neurosurgeon (Kang et al., 2017) and DDNN (Teerapittayanan et al., 2017) introduced automated frameworks to dynamically distribute DNN workloads between mobile devices and cloud servers, optimizing for factors such as network latency and energy efficiency. Meanwhile, BranchyNet (Teerapittayanan et al., 2016) incorporated early-exit branches within DNN architectures, enabling adaptive inference based on input complexity to further reduce end-to-end processing latency. To the best of our knowledge, DSD represents the first split computing framework specifically designed for speculative decoding.

2.3 Simulation Framework for Large Models

As LLMs continue to increase in both complexity and deployment scale, their performance evaluation increasingly depends on simulation frameworks that capture the interplay among workload characteristics, system architecture, and resource constraints. These frameworks operate across multiple levels of abstraction, ranging from GPU microarchitecture to datacenter-scale distributed systems. In the following, we review representative simulators at each level.

GPU-Level Simulators. At the hardware level, GPU simulators enable cycle-accurate modeling of GPU architectures for diverse workloads. GPGPU-Sim (Bakhoda et al., 2009) offers detailed simulation of modern NVIDIA GPUs executing CUDA applications, with support for Tensor Cores, dynamic parallelism, and integrated energy estimation via GPUWattch. Building upon this foundation, Accel-Sim (Khairy et al., 2020) extends the framework with an extensible, trace-driven design that directly consumes native SASS instructions, substantially reducing the effort needed to model new GPU architectures while preserving validation

accuracy within 20% of real hardware measurements.

LLM Inference Simulators. The emergence of large language model (LLM) serving systems has driven the development of specialized simulation frameworks that capture the unique dynamics of autoregressive inference. VIDUR (Agrawal et al., 2024b), the first large-scale simulator for LLM inference, models operator performance using empirical profiling and predictive modeling, achieving an average latency estimation error of approximately 9%. Building on this direction, LLMServingSim (Li et al., 2024) advances hardware-software co-simulation by extending ASTRA-sim to jointly model LLM-specific accelerators and serving system software. It integrates GPU kernel simulation with system-level request scheduling and batching, providing up to 91.5 \times faster simulation than prior accelerator simulators while maintaining an error below 14.7% compared to real GPU-based serving systems. Complementing these efforts, ReaLLM (Peng et al., 2025) introduces a trace-driven framework with a built-in generator that leverages production workloads from the Azure LLM Inference Dataset. Through linear interpolation-based latency prediction for matrix multiplication, ReaLLM accelerates simulation while maintaining low error rates, enabling rapid evaluation of service-level objectives across diverse system configurations. DSD-Sim leverages the VIDUR framework as its foundation for modeling the performance of distributed SD, yet it maintains compatibility with all previously discussed LLM simulation frameworks.

Distributed Neural Network Simulators. Numerous efforts have been made to simulate the behavior of distributed training and inference. ASTRA-sim (Akbari et al., 2020; Chang et al., 2023) models the end-to-end software-hardware stack, encompassing workload scheduling, collective communication algorithms, and hierarchical network topologies. Its second version extends this capability by supporting arbitrary model-parallel strategies through graph-based implementations and by providing analytical performance estimates for multi-dimensional heterogeneous topologies, thereby enabling scalable simulation-based exploration of distributed training platforms.

Advanced LLM Serving Frameworks. Recent studies highlight the critical role of specialized architectures and scheduling policies in optimizing LLM serving. Dist-Serve (Zhong et al., 2024) disaggregates prefill and decoding stages across distinct GPUs to eliminate phase interference, achieving up to 7.4 \times higher goodput under stringent Service-Level Objectives (SLO) constraints. Sarathi (Agrawal et al., 2023) proposes chunked prefill, which partitions long input prompts into fixed-size chunks to enable decode-maximal batching, improving decode throughput by as much as 10 \times for LLaMA-13B. Build-

ing on this, Sarathi-Serve (Agrawal et al., 2024a) introduces stall-free scheduling mechanisms to further enhance latency control during serving.

While existing frameworks effectively model single-node LLM inference or distributed training, they fall short in addressing the unique challenges of distributed SD. Unlike conventional serving, where each request is executed independently on a single device, distributed SD introduces cross-device dependencies between draft generation and target verification, with performance tightly coupled to network round-trip time, token acceptance rates, and dynamic speculation window size. Furthermore, the interaction among prefill queueing, speculation window size, and target model utilization leads to complex performance bottlenecks that cannot be captured by single-node simulators such as VIDUR or system-agnostic frameworks like ASTRA-sim. These limitations motivate the development of a specialized simulation framework.

3 SIMULATION FRAMEWORK FOR DISTRIBUTED SPECULATIVE DECODING

As discussed in Section 2.3, due to the lack of an existing simulation framework for distributed SD, we introduce *DSD-Sim*, a request-level discrete-event simulator tailored for distributed SD. Given a deployment description and workload traces, DSD-Sim models routing, batching, speculation/verification iterations, and network effects to produce per-request and system-level SLO metrics. Figure 2 summarizes the data flow.

3.1 Overview

As shown in Figure 2, DSD-Sim comprises four primary components: a *configuration parser*, a core *DSD scheduler*, a *hardware performance modeling engine* that integrates validated single-node predictors, and a *performance analyzer* that records both static and dynamic system metrics.

The configuration parser ingests system specifications from a YAML file that defines device types (e.g., model, hardware), network links (e.g., RTT, jitter), and runtime policies. An `auto_topology` pass expands this high-level specification into explicit draft and target device pools with fully defined network connections, enabling rapid system configuration and large-scale parameter sweeps.

The core DSD scheduler leverages SimPy to model draft and target servers as concurrent processes, each with explicit queues for batch formation and request scheduling. Network links are represented as delay elements associated with send and receive events, parameterized by configurable RTT and jitter. This request-level abstraction allows fast exploration of large design spaces while preserving accuracy in modeling system-level queueing and network dynamics.

To accurately capture per-kernel execution latency without reimplementing low-level GPU models, DSD-Sim incorporates a hardware performance modeling engine based on VIDUR (Agrawal et al., 2024b). VIDUR provides validated latency predictors for both *prefill* and *decode* operations, grounded in empirical measurements from contemporary GPU architectures.

During simulation, the DSD scheduler communicates with VIDUR’s single-node predictors via a unified API, `predict(op, shape, hardware)`, which allows querying inference latency for arbitrary batch compositions across heterogeneous devices. This design enables DSD-Sim to retain VIDUR’s single-node accuracy while extending it to distributed inference scenarios. To further adapt VIDUR to distributed speculative decoding workloads, we expand its profiling dataset with measurements from NVIDIA A40 GPUs and edge-oriented LLMs such as Qwen-7B (Bai et al., 2023) and Llama2-7B (Touvron et al., 2023a). These additions enable DSD-Sim to model latency behaviors under heterogeneous and resource-constrained environments, improving its predictive accuracy for real-world edge deployments.

3.2 Workloads and Trace Model

DSD-Sim is driven by workload traces that embed request parameters and ground-truth speculation outcomes captured from hardware.

Trace Schema Traces are derived from representative benchmarks spanning diverse LLM tasks such as GSM8K (Cobbe et al., 2021), CNN/DailyMail (Hermann et al., 2015), and HumanEval (Chen et al., 2021). Each record specifies the parameters needed to drive a simulation run. Crucially, the embedded acceptance sequence is captured from real hardware profiling runs, which faithfully reproduce speculation behavior for a given draft–target pair without relying on probabilistic acceptance models. Table 1 lists the key fields in a trace record. The selected benchmarks capture a wide range of LLM inference behaviors, including reasoning-intensive, summarization, and code-generation tasks, covering different output-to-input ratios and speculation acceptance dynamics. This diversity ensures that DSD-Sim operates under representative real-world LLM serving conditions.

Arrival Process Request arrivals can operate in two modes: (i) trace-driven, where timestamps are replayed from a captured workload, and (ii) synthetic, where arrivals follow a Poisson process with a specified rate to emulate stochastic load. In the synthetic mode, arrivals are generated globally and uniformly distributed across drafter devices.

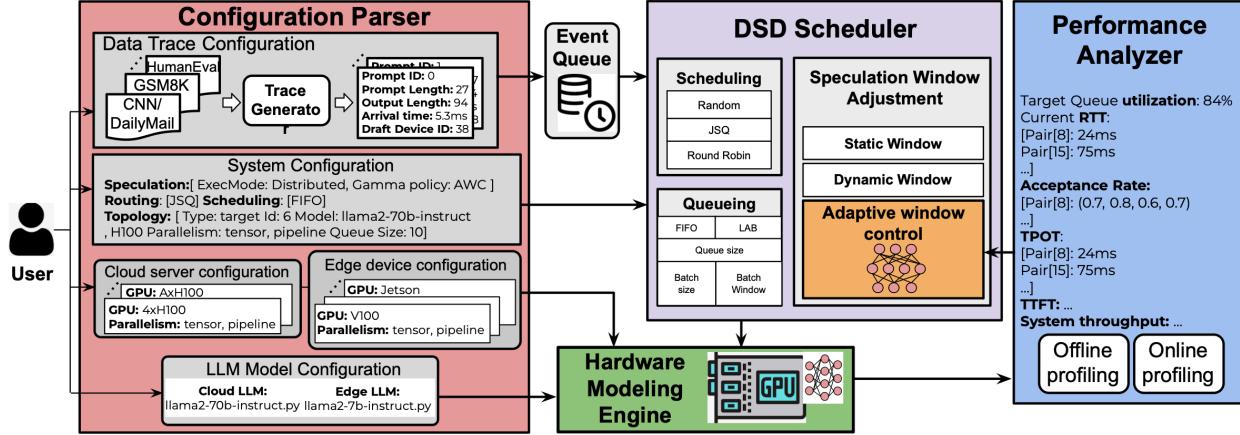


Figure 2. The configuration parser ingests YAML configuration files and workload traces, routing requests through the DSD scheduler. The scheduler then coordinates the hardware modeling engine to generate detailed system performance outputs, which are subsequently processed by the performance analyzer for SLO evaluation.

Table 1. Key fields of a workload trace record.

Field Name	Example Value
prompt_length	27
output_length	94
acceptance_seq	[1, 0, 1, ...]
arrival_time_ms	5.3
drafter_id	38

3.3 Execution Semantics of DSD-Sim

Each request follows a lifecycle that includes routing, batching, and iterative speculation until the generation process completes. DSD-Sim supports two execution modes: (i) **Fused mode**, in which both draft and target models are co-located on the same server. The entire speculation loop executes locally, eliminating network overhead between the drafter and verifier. (ii) **Distributed mode**, where a draft device proposes a window of tokens (speculation window size $\gamma > 1$), which are then transmitted to a remote target device for parallel verification.

For each arriving request, the simulation progresses through **Routing**, **Batching**, **Speculation**, and **Verification** stages. During *Routing*, the request is directed to a target cluster according to the active scheduling policy. In *Batching*, it is placed into a queue and grouped with other requests for prefill or decode operations. During *Speculation*, the active γ policy determines the window size γ and execution mode; in distributed mode, a drafter device generates γ speculative tokens and transmits them to the target for verification. In *Verification*, the target server compares these tokens with its own predictions. Accepted tokens advance the generated

sequence, while mismatched tokens are corrected using the target's prediction, and any remaining speculative tokens are discarded. This iterative process continues until the desired output length is reached or an end-of-sequence token is produced.

3.4 DSD Scheduling Policies

The execution lifecycle is governed by three families of pluggable policies. Each policy operates on a read-only snapshot of recent system performance metrics, such as queue depth, round-trip time (RTT), time-per-output-token (TPOT), and acceptance rate.

Request Routing Policy These policies determine how incoming requests are distributed across clusters, supporting a variety of load-balancing strategies in multi-cluster deployments. Common implementations include random selection, round-robin scheduling, and the Join-the-Shortest-Queue (JSQ) policy.

Batching Policy These policies govern how requests are grouped for prefill and decode operations. They are controlled by configurable parameters for batch size and batching window, and can optionally enable advanced strategies like continuous batching or chunked prefills.

Window Size Policy These policies dynamically select the appropriate speculation window size (γ) and execution mode (fused vs. distributed) for each iteration. The simulator provides several such policies: *Static window*, which fixes γ to a constant value; *Dynamic window*, which uses threshold-based heuristics to adjust γ based on the recent acceptance rate; and our primary learned policy, *Adaptive*

Window Control (AWC), which is described in detail in Section 4.

3.5 Performance Metrics

To enable policy optimization and offline analysis, DSD-Sim captures detailed performance data for processing by the Performance Analyzer. This data serves as ground truth for evaluating system performance against SLOs and as training input for AWC (Section 4). Specifically, we collect two categories of metrics:

- **Per-Request Metrics:** Time-to-first-token (TTFT), time-per-output-token (TPOT), end-to-end latency, acceptance ratios, routing decisions, and the sequence of window size (γ) decisions.
- **System-Level Metrics:** Overall system throughput, target device utilization, and aggregate network queueing delays.

All measurements are emitted in a structured JSON format to facilitate both online policy adaptation and offline analysis.

4 ADAPTIVE WINDOW CONTROL

The choice of the speculation window size γ is crucial for the overall throughput of the distributed framework. If γ is set too large, it increases the latency of the draft model and raises the likelihood of verification failures. Conversely, setting γ too small results in overly frequent verification steps, significantly increasing both the computational load on the cloud LLM and the communication overhead required for verification tasks.

In real-world deployments, factors such as network latency, token acceptance rate, and device utilization often fluctuate over time, making it challenging for fixed or heuristic-based γ policies to consistently balance efficiency and responsiveness. A static speculation window can easily cause excessive rollback overhead under adverse network conditions or lead to severely underutilized hardware when more aggressive speculation would be beneficial. Moreover, it fails to adapt to dynamic system-level conditions in large-scale distributed environments, such as network congestion, queue buildup, and evolving workload characteristics.

To address this issue, Adaptive Window Control (AWC) dynamically adjusts the window size γ to achieve superior overall speedup in DSD. By modeling the correlations among system load, communication delay, and token acceptance dynamics, AWC offers a principled, data-driven approach to maintaining high speedup and stable performance across heterogeneous hardware and network conditions.

4.1 Data-Driven Window Predictor Design

During runtime, AWC utilizes a window control deep neural network (WC-DNN) that receives input from the performance analyzer. The analyzer encodes both current and historical system states into a feature vector, which the DNN uses to predict the optimal value of γ . Specifically, this feature vector includes the following information:

- **Queue Depth Utilization (q_{depth}):** Recent utilization rate of the target queue, reflecting the current level of congestion.
- **Acceptance Rate (α_{recent}):** Recent token acceptance ratio from the target model, reflecting the quality and reliability of the draft model's generated tokens.
- **Per-Link RTT Statistics ($\text{RTT}_{\text{recent}}$):** Recent round-trip latency of the network connection, used to determine the viability of distributed execution.
- **TPOT Statistics ($\text{TPOT}_{\text{recent}}$):** Recent time per output token, representing the processing efficiency of the target model.
- **Prior Speculation Window Size (γ_{prev}):** Speculation window size from the previous iteration, providing temporal context for adaptive adjustment.

The WC-DNN processes this input to produce the optimal prediction for γ . The compact size of the input feature vector allows the WC-DNN to operate with minimal computational and memory overhead, thereby ensuring fast and efficient adaptation during runtime.

4.2 Dataset Generation

Training the WC-DNN requires labeled data that maps system feature contexts to their optimal window size configurations. To create this dataset, we perform exhaustive sweeps of the window size under various system conditions. For each experimental setup, defined by a workload trace, network configuration, and hardware deployment, the simulator runs across all combinations of window sizes (from 2 to 12) as well as the fused execution mode.

During each sweep, the system records a set of data that includes the feature vector, policy output, and performance metrics such as TTFT and TPOT. After gathering results from more than 2,000 scenarios, we construct training labels by selecting the configuration that minimizes a weighted objective function reflecting the SLO priorities.

4.3 Model Architecture and Training

The WC-DNN adopts a residual multi-layer perceptron (MLP) architecture that predicts the speculation window

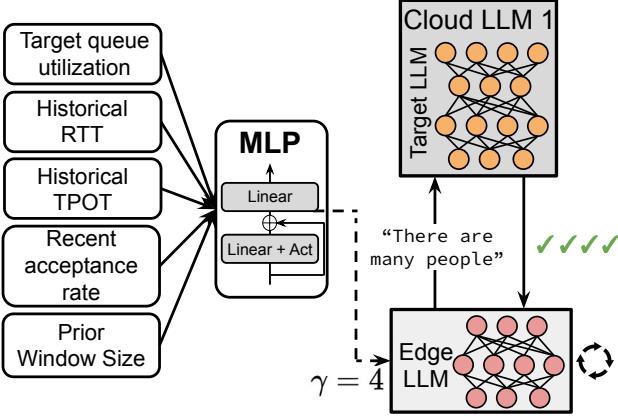


Figure 3. The WC-DNN architecture takes five input features and generates the optimal prediction for the speculation window size. The edge LLM then adjusts its window size based on this prediction and coordinates with the cloud LLM for verification.

size γ as a continuous value, as shown in Figure 3. The MLP takes a five-dimensional feature vector as input and employs a residual structure with two computational blocks and SiLU activation functions to generate a stable scalar prediction for the window size. The network is trained using supervised regression with an L_1 loss, and we use the AdamW optimizer for 100 epochs, achieving consistently high predictive accuracy on the validation set.

4.4 Stable Execution of Window Size Prediction

In practice, directly executing the trained WC-DNN can lead to fluctuations in the predicted window size γ because of variations in system metrics, which in turn cause instability in overall system performance. To address this issue, we introduce three stabilization techniques:

1. **Clamping:** Window size predictions are clipped to a configured range (e.g., [1, 12]).
2. **Exponential Smoothing:** An Exponential Moving Average (EMA) with a smoothing factor $\alpha = 0.4$ is applied to the predicted window size across iterations to dampen high-frequency oscillations.
3. **Hysteresis for Mode Switching:** A “sticky” policy prevents rapid toggling between fused and distributed modes. When the system is in distributed mode, the smoothed prediction must remain near $\gamma = 1$ for at least k consecutive steps (typically $k = 2$) before a switch to fused mode is permitted.

After applying the stabilization techniques, the predicted window size is quantized to the nearest integer within the predefined range (e.g., [1, 12]) and then used in the next iteration of SD. The smoothing state is maintained per

draft-target pair so each connection follows its own trajectory, but shared features ensure that decisions still reflect aggregate system conditions.

Crucially, if the window size predicted by the WC-DNN is less than or equal to one ($\gamma \leq 1$), the system switches to **Fused Mode**. In this state, the cloud LLM generates all tokens directly, bypassing the draft model. This situation typically arises when the edge device operates very slowly or when network conditions are severely congested.

5 EVALUATION

We evaluate DSD-Sim across two primary dimensions: (1) *System Performance and Scalability*, by testing diverse configurations and workload compositions across heterogeneous Edge–Cloud setups; and (2) *Policy Effectiveness*, by analyzing how the proposed AWC strategy performs relative to conventional scheduling and batching baselines.

- **Calibration:** How closely do VIDUR’s latency predictions and our RTT assumptions match real hardware measurements?
- **Scalability:** How does distributed speculative decoding behave as the system scales in heterogeneity and the number of participating nodes?
- **Policy Efficiency:** How effectively does AWC enhance resource utilization and reduce the overall LLM execution latency compared to other baseline policies?

All experiments are conducted using workload traces derived from GSM8K (Cobbe et al., 2021), CNN/DailyMail (Hermann et al., 2015), and HumanEval (Chen et al., 2021) benchmarks. Specifically, we use Qwen-7B (Bai et al., 2023) and Llama2-7B (Touvron et al., 2023a) running on NVIDIA A40 GPUs (NVI, 2021) to model edge-side inference, while Llama2-70B (Touvron et al., 2023a) and Qwen-72B (Yang et al., 2024) are deployed on A100 (NVI, 2020) and H100 GPUs (NVI, 2022), respectively, to represent cloud-scale execution. Each measurement is repeated across multiple random seeds, and the reported results represent the mean values to reduce the influence of stochastic variations.

5.1 Calibration and Validation

DSD-Sim is composed of several key components: a configuration parser that loads traces and system settings, a deterministic discrete-event scheduler that applies the specified routing, queueing, and window policies, a hardware modeling engine that estimates GPU inference latency, and a performance analyzer that records and reports system-level metrics. Only the hardware modeling engine, which

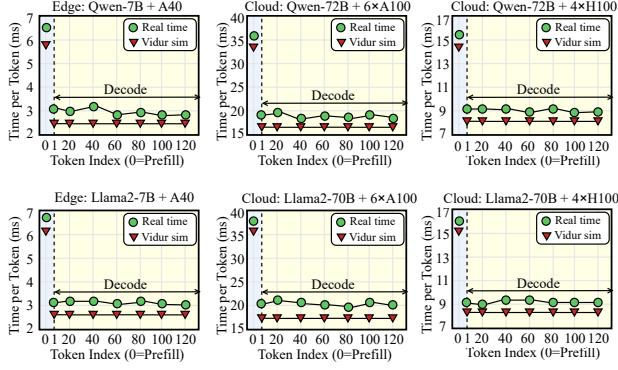


Figure 4. GPU-level calibration of predicted vs. actual inference latencies for prefill and decode across Qwen-7B, Qwen-72B, Llama2-7B, and Llama2-70B on A40, A100, and H100 GPUs. Error bars indicate standard deviation over 100 requests.

is built on VIDUR, and the assumed edge–cloud RTT introduce modeling uncertainty; the parser and scheduler simply replay traces and apply policies without any tunable parameters. Therefore, we validate VIDUR’s prefill and decode latency predictions against real hardware measurements and determine RTT values using real measurements from prior latency studies. With these components calibrated, the simulator can reasonably estimate system behavior while maintaining a controlled and reproducible evaluation setting.

GPU-Level Calibration Figure 4 presents a comparison between DSD-Sim’s predicted inference latencies and empirical measurements obtained from real GPU deployments across different model sizes and batch configurations. In particular, we evaluate simulation fidelity using multiple large language models, including Qwen-7B, Qwen-72B (Team et al., 2024), Llama2-7B, and Llama2-70B (Touvron et al., 2023b), executed on diverse real GPU platforms such as NVIDIA A40, A100, and H100. To ensure consistency in workload patterns, all models are benchmarked on prompts derived from the GSM8K dataset, which provides realistic reasoning-oriented text sequences representative of modern LLM inference workloads. In parallel, the hardware modeling engine is executed to estimate the corresponding simulated latencies, enabling direct calibration and cross-validation against real-world hardware performance. Specifically, we adopt VIDUR (Agrawal et al., 2024b) to simulate GPU hardware performance. For smaller models such as Qwen-7B and Llama2-7B, a single GPU is sufficient to accommodate the entire model. In contrast, larger models such as Qwen-72B and Llama2-70B require multi-GPU execution, where model parallelism is employed to distribute weights and computation across multiple devices.

The simulator integrates VIDUR’s performance predictors and achieves a mean absolute error of 7.4% for prefill la-

tency and 5.2% for decode latency across all evaluated configurations. This level of accuracy validates the simulator’s reliability in modeling single-node LLM inference behavior.

It is worth noting that VIDUR’s simulated latencies are consistently lower than empirical measurements across all model and GPU setups. This systematic deviation stems from the fact that VIDUR focuses exclusively on modeling MLP and Attention kernel execution times, omitting NCCL communication overheads (ncc, 2018) and other non-kernel computations that occur in real hardware environments.

Network Calibration We model the edge–to–cloud communication link using two representative round-trip time (RTT) settings. According to public Azure latency reports, the RTT to nearby East-US regions is typically below 20 ms (azu, 2024), while independent measurement studies on Azure and AWS observe values ranging from 10 to 30 ms for clients located in the same geographic area as the serving region (Palumbo et al., 2020). Guided by these observations, we evaluate network performance under two conditions—10 ms representing the typical case and 30 ms representing the upper bound of common operating scenarios. DSD-Sim reports performance metrics for both networking settings.

5.2 System Performance Analysis

In this section, we evaluate the capacity of DSD-Sim to model distributed speculative decoding across a range of hardware configurations, network conditions, and execution modes.

Performance Across System Configurations We define a large-scale heterogeneous cluster comprising a **Cloud Pool** and an **Edge Pool**. The Cloud Pool consists of 20 servers hosting large models (LLaMA2-70B, LLaMA3-70B, Qwen-72B) distributed across 4×A100, 4×H100, and 4×A6000 configurations. The Edge Pool contains 600 GPUs (300 A40s and 300 V100s) evenly serving LLaMA-2-7B, Qwen-7B, and LLaMA-3.1-8B draft models. All experiments run over a 10 ms RTT network connection.

The scheduler exposes three knobs: **Routing** (Random vs. JSQ); **Queueing** (FIFO vs. Length-Aware Batching, or LAB, which groups requests of similar lengths to minimize padding); and **Window Control** (Static $\gamma = 4$, Dynamic γ , and AWC). Figure 5 compares five policy stacks: Default (Random + FIFO + Static γ), Setting 1 (JSQ + FIFO + Static γ), Setting 2 (JSQ + LAB + Static γ), Setting 3 (JSQ + LAB + Dynamic γ), and Setting 4 (JSQ + LAB + AWC).

Accumulating these policies yields steady improvements in both throughput and latency. Across GSM8K, for example, throughput climbs from 25.1 to 28.1 requests/s as we add all four components, while TTFT drops from 351 to 345 ms and TPOT from 45 to 37 ms. CNN/DailyMail and

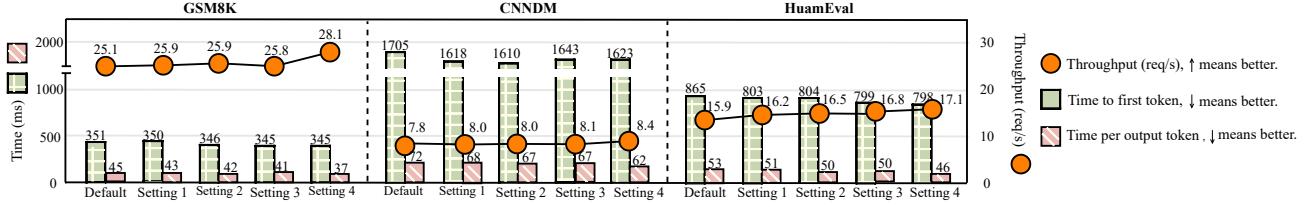


Figure 5. End-to-end SLOs and throughput for policy stacks. Default: Random routing + FIFO queueing + Static γ . Setting 1: JSQ + FIFO + Static γ . Setting 2: JSQ + Length-Aware Batching (LAB) + Static γ . Setting 3: JSQ + Length-Aware Batching + Dynamic γ . Setting 4: JSQ + Length-Aware Batching + AWC.

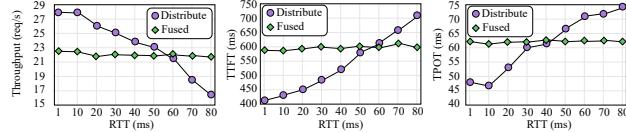


Figure 6. Throughput, TTFT, and TPOT for distributed (purple) vs. fused (green) execution as RTT increases. Distributed excels at low RTT but degrades once network delay dominates; fused remains steady because work stays local.

HumanEval exhibit similar trends, with AWC providing the major latency gain. The overall results demonstrate that coordinating routing, batching, and adaptive window control jointly offers the optimal throughput-latency trade-off.

Distributed vs. Cloud-Only Execution We next analyze how network RTT affects the performance of distributed speculative decoding. Figure 6 varies RTT while keeping all other system configurations fixed. At low RTT values, distributed execution outperforms cloud-only mode because edge drafts are generated concurrently with cloud verification, improving throughput. However, as RTT increases, the communication overhead of each speculative iteration grows, causing noticeable degradation. In contrast, cloud-only mode (where the target handles both drafting and verification locally) remains unaffected by network latency but suffers from lower parallelism at low RTTs. The performance crossover observed around 50–60 ms explicitly highlights the trade-off between compute offloading and network overhead, validating the need for AWC’s adaptive mode switching.

AWC vs. Baseline Policies We next evaluate AWC against two baselines: *Static Window* (fixed $\gamma = 4$) and *Dynamic Window* (increments γ when the recent acceptance rate exceeds 0.75, and decrements when it falls below 0.25). Table 2 presents results across four system configurations, combining 20 targets with either 600 or 1000 drafts and network RTTs of 10 ms or 30 ms, evaluated on three datasets: GSM8K, CNN/DailyMail, and HumanEval. AWC delivers the best throughput in all 12 cases, improving over the static

baseline by 3–10% (e.g., +9.7% on GSM8K with 600 drafts and +4.1% on CNNDM with 1000 drafts) and outperforming the dynamic baseline in 11 out of 12 scenarios.

Latency metrics closely mirror these overall trends. TTFT remains within 0.5–4% of the best baseline, and TPOT consistently drops by 6–10% across the evaluated workload set (400 GSM8K, 400 CNN/DailyMail, and 100 HumanEval prompts). These results clearly show that a learned controller can simultaneously raise throughput and lower latency across heterogeneous conditions without manual threshold tuning.

5.3 Ablation Studies

In this section, we perform systematic ablation studies to isolate and quantify the effects of key design choices in scheduling, queueing, and batching policies. While the previous section demonstrated the aggregate benefits of the full DSD stack, identifying the individual contribution of each component is critical for system optimization. The complex interaction between load balancing, batch formation, and window sizing often obscures the specific source of performance gains. Therefore, we decouple these mechanisms to analyze how routing decisions impact server utilization and how batching strategies mitigate head-of-line blocking.

Request Routing Strategies Figure 8 evaluates system performance by varying the number of draft models from 0.4k to 2.0k while comparing three routing strategies: Random, Round-Robin (RR), and Join-Shortest-Queue (JSQ), as described in Section 3.4.

JSQ achieves the best performance when system resources are not fully utilized, maintaining TPOT values 5–20 ms lower across all traces by routing requests to less busy target servers. However, as the system approaches its maximum capacity, the fastest server remains saturated, causing subsequent requests already assigned to it to suffer from head-of-line blocking. As a result, the TPOT curve eventually rises above that of Round-Robin.

Although RR is not load-aware, it distributes user requests more evenly across servers. The throughput trends in Fig-

Table 2. Adaptive window control versus baseline γ policies. Values are averaged over three runs; parentheses indicate improvement relative to the Static baseline (positive for higher throughput, negative for lower latency).

Config 1: 20T/600D, 10ms RTT			Config 2: 20T/1000D, 10ms RTT			Config 3: 20T/600D, 30ms RTT			Config 4: 20T/1000D, 30ms RTT			
Static	Simple	AWC	Static	Simple	AWC	Static	Simple	AWC	Static	Simple	AWC	
Throughput (requests/s) \uparrow												
GSMK	25.8	26.1	28.3 (+9.7%)	30.6	30.7	31.5 (+3.0%)	17.8	18.9	18.5 (+3.9%)	21.4	21.8	22.3 (+4.2%)
HumanEval	16.2	16.7	17.2 (+6.1%)	17.4	17.6	18.2 (+4.6%)	11.2	11.0	11.3 (+0.9%)	12.8	13.1	13.2 (+3.1%)
CNNMD	8.0	8.25	8.4 (+5.0%)	12.1	12.2	12.6 (+4.1%)	6.0	6.2	6.2 (+3.3%)	9.3	9.5	9.7 (+4.3%)
TTFT (ms) \downarrow												
GSM8K	352	334	347 (-1.4%)	368	354	362 (-1.6%)	462	457	443 (-4.1%)	478	465	463 (-3.1%)
HumanEval	803	792	799 (-0.5%)	2102	2178	2038 (-3.0%)	1138	1125	1120 (-1.6%)	2578	2520	2497 (-3.1%)
CNNMD	1602	1625	1569 (-2.1%)	2412	2408	2352 (-2.5%)	2024	1993	2006 (-0.9%)	2986	2934	2925 (-2.0%)
TPOT (ms) \downarrow												
GSM8K	40.1	38.9	36.2 (-10.7%)	72.2	69.7	67.4 (-7.1%)	62.3	62.9	58.3 (-6.8%)	99.4	98.1	91.5 (-8.6%)
HumanEval	50.8	48.6	46.3 (-9.7%)	75.4	73.4	70.5 (-6.9%)	79.2	78.1	73.6 (-7.0%)	102.7	100.2	96.8 (-6.0%)
CNNMD	67.4	66.9	62.8 (-6.8%)	82.1	82.6	76.8 (-6.5%)	92.4	88.6	86.7 (-6.2%)	108.3	108.5	98.2 (-10.2%)

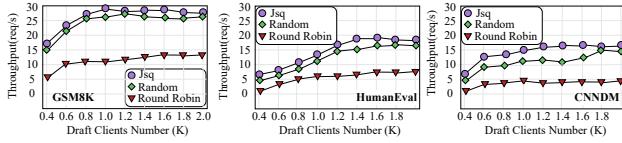


Figure 7. Throughput vs. the number of draft clients for the same routing policies. JSQ consistently delivers the best throughput up to ~ 1 k drafts, but gradually saturates thereafter.

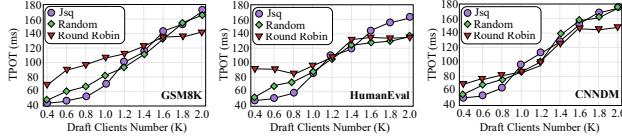


Figure 8. TPOT versus number of draft clients for three routing policies (GSM8K, HumanEval, CNNMD).

ure 7 reflect a similar pattern: JSQ delivers the highest scalability up to around one thousand draft models but then plateaus as resources become saturated, while RR continues to improve and eventually catches up. These results indicate that adaptive scheduling strategies, which respond to changing system conditions, can outperform static policies and maintain stable performance across varying load levels.

Queueing and Batching Policies We compare simple FIFO dispatch with a length-aware batching (LAB) policy. LAB takes the head-of-line request and batches it with other requests whose lengths closely match the head-of-line request. It is widely adopted (e.g., ORCA (Yu et al., 2022), Sarathi (Agrawal et al., 2023)) and serves as the strongest length-aware baseline. Figure 9 illustrates the latency benefit of this strategy. LAB consistently achieves lower TPOT compared to FIFO (reducing latency by 1–2 ms) across all workloads. By grouping requests with similar execution lengths, LAB minimizes the padding overhead within a batch, effectively mitigating head-of-line blocking where

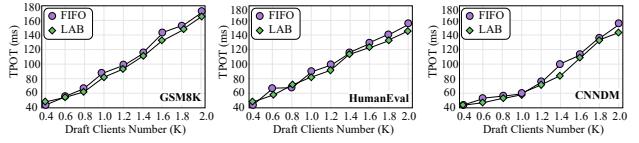


Figure 9. FIFO versus length-aware batching (LAB) performance comparison across diverse workloads. LAB groups similar-length jobs, improving TPOT under moderate-to-high load conditions.

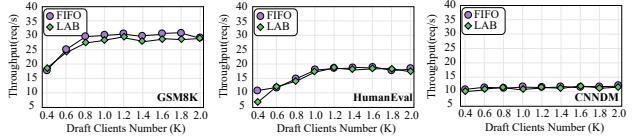


Figure 10. FIFO versus length-aware batching (LAB) across workloads. LAB groups similar-length jobs, improving throughput.

short requests wait for long ones. In terms of scalability, Figure 10 demonstrates that both policies reach similar throughput ceiling. As the number of draft clients increases beyond 1k, the system saturates, and throughput plateaus for both FIFO and LAB. This confirms that while LAB optimizes execution efficiency (improving latency), the maximum aggregate throughput is determined by the underlying compute capacity rather than the queueing order.

6 CONCLUSION

We presented DSD, a distributed speculative decoding framework for efficient large model serving across edge–cloud environments. Built upon the DSD-Sim simulator, DSD models realistic network, batching, and scheduling dynamics to guide performance optimization. We further introduced AWC, a data-driven policy that dynamically adjusts the speculation window for optimal throughput and stability. Experiments show that DSD greatly improves latency and scalability over existing SD methods.

REFERENCES

- Nvidia collective communications library (nccl), 2018. <https://developer.nvidia.com/nccl>.
- Pysyft, 2020. <https://github.com/OpenMined/PySyft>.
- Splitnn, 2020. <https://blog.openmined.org/tag/splitnn/>.
- Azure network round-trip latency statistics. <https://learn.microsoft.com/en-us/azure/networking/azure-network-latency>, 2024.
- Agrawal, A., Panwar, A., Mohan, J., Kwatra, N., Gulavani, B. S., and Ramjee, R. Sarathi: Efficient llm inference by piggybacking decodes with chunked prefills. *arXiv preprint arXiv:2308.16369*, 2023.
- Agrawal, A., Kedia, N., Panwar, A., Mohan, J., Kwatra, N., Gulavani, B. S., Tumanov, A., and Ramjee, R. Taming throughput-latency tradeoff in llm inference with sarathi-serve. In *Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2024a.
- Agrawal, A., Kedia, N., Panwar, A., Mohan, J., Kwatra, N., Gulavani, B. S., Tumanov, A., and Ramjee, R. Vidur: A large-scale simulation framework for llm inference. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 6, 2024b.
- Akbari, S., Zhao, T., Ankit, A., Ogras, U. Y., Dasu, A., Kundu, S., Panda, P., Jain, A., Parthasarathy, S., Swaminathan, K., et al. Astra-sim: Enabling sw/hw co-design exploration for distributed dl training platforms. In *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2020.
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. Qwen technical report, 2023. URL <https://arxiv.org/abs/2309.16609>.
- Bakhoda, A., Yuan, G. L., Fung, W. W., Wong, H., and Aamodt, T. M. Gpgpu-sim: A cycle-level gpu simulator. In *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2009.
- Chang, W., Saba, T., Samtani, S., Li, J., Samsi, S., Gadepally, V., Hoefer, T., Jain, T., et al. Astra-sim2.0: Modeling hierarchical networks and disaggregated systems for large-model training at scale. In *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2023.
- Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., and Jumper, J. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillett, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Ding, A., Hass, A., Chan, M., Sehatbakhsh, N., and Zonouz, S. Resource-aware dnn partitioning for privacy-sensitive edge-cloud systems. In *International Conference on Neural Information Processing*, pp. 188–201. Springer, 2023.
- Dong, X., Zhang, S. Q., Li, A., and Kung, H. Spherefed: Hyperspherical federated learning. In *European Conference on Computer Vision*, pp. 165–184. Springer, 2022.
- Feltin, T., Marchó, L., Cordero-Fuertes, J.-A., Brockners, F., and Clausen, T. H. Dnn partitioning for inference throughput acceleration at the edge. *IEEE Access*, 11: 52236–52249, 2023.
- Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., and Zhang, J. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*, 2023.
- Hauswald, J., Manville, T., Zheng, Q., Dreslinski, R., Chakrabarti, C., and Mudge, T. A hybrid approach to offloading mobile image classification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8375–8379. IEEE, 2014.

- Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pp. 1693–1701, 2015.
- Kang, W., Chung, S., Kim, J. Y., Lee, Y., Lee, K., Lee, J., Shin, K. G., and Chwa, H. S. Dnn-sam: Split-and-merge dnn execution for real-time object detection. In *2022 IEEE 28th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 160–172. IEEE, 2022.
- Kang, Y., Hauswald, J., Gao, C., Rovinski, A., Mudge, T., Mars, J., and Tang, L. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News*, 45(1):615–629, 2017.
- Karjee, J., Naik, P., Anand, K., and Bhargav, V. N. Split computing: Dnn inference partition with load balancing in iot-edge platform for beyond 5g. *Measurement: Sensors*, 23:100409, 2022.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103: 102274, 2023.
- Khairy, M., Shen, Z., Aamodt, T. M., and Rogers, T. G. Accel-sim: An extensible simulation framework for validated gpu modeling. In *Proceedings of the 47th International Symposium on Computer Architecture (ISCA)*, 2020.
- Lee, J., Lee, H., and Choi, W. Wireless channel adaptive dnn split inference for resource-constrained edge devices. *IEEE Communications Letters*, 2023.
- Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.
- Li, J., Lee, Y., Jang, J., Kim, Y., Kwon, J. H., and Kim, J. H. Llmservingsim: A hw/sw co-simulation infrastructure for llm inference serving at scale. In *Proceedings of the IEEE International Symposium on Workload Characterization (IISWC)*, 2024.
- Lin, J., Li, M., Zhang, S. Q., and Leon-Garcia, A. Murmuration: On-the-fly dnn adaptation for slo-aware distributed inference in dynamic edge environments. In *Proceedings of the 53rd International Conference on Parallel Processing*, pp. 792–801, 2024.
- Liu, T., Li, Y., Lv, Q., Liu, K., Zhu, J., Hu, W., and Sun, X. Pearl: Parallel speculative decoding with adaptive draft length. *arXiv preprint arXiv:2408.11850*, 2024.
- Luo, D., Yu, T., Wu, Y., Wu, H., Wang, T., and Zhang, W. Split: Qos-aware dnn inference on shared gpu via evenly-sized model splitting. In *Proceedings of the 52nd International Conference on Parallel Processing*, pp. 605–614, 2023.
- Matsubara, Y., Baidya, S., Callegaro, D., Levorato, M., and Singh, S. Distilled split deep neural networks for edge-assisted real-time systems. In *Proceedings of the 2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges*, pp. 21–26, 2019.
- Matsubara, Y., Levorato, M., and Restuccia, F. Split computing and early exiting for deep learning applications: Survey and research challenges. *ACM Computing Surveys*, 55(5):1–30, 2022.
- Mubark, W. H., Kasula, J. G., and Sarwar Uddin, M. Y. Asap: Asynchronous split inference for accelerated dnn execution. In *Proceedings of the 25th International Conference on Distributed Computing and Networking*, pp. 32–44, 2024.
- NVIDIA A100 Tensor Core GPU Architecture. NVIDIA Corporation, 2020. Available at <https://www.nvidia.com/en-us/data-center/a100/>.
- NVIDIA A40 GPU Architecture. NVIDIA Corporation, 2021. Available at <https://www.nvidia.com/en-us/data-center/a40/>.
- NVIDIA H100 Tensor Core GPU Architecture. NVIDIA Corporation, 2022. Available at <https://www.nvidia.com/en-us/data-center/h100/>.
- OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Palumbo, F., Aceto, G., Botta, A., Ciuonzo, D., Persico, V., and Pescapè, A. Characterization and analysis of cloud-to-user latency: the case of azure and aws. *Computer Networks*, 184:107693, 2020.
- Peng, Y. et al. Reallm: A trace-driven framework for rapid simulation of llm inference systems. In *Proceedings of the 36th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, 2025.
- Stern, M., Shazeer, N., and Uszkoreit, J. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.
- Team, Q. et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

- Teerapittayanon, S., McDanel, B., and Kung, H.-T. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pp. 2464–2469. IEEE, 2016.
- Teerapittayanon, S., McDanel, B., and Kung, H.-T. Distributed deep neural networks over the cloud, the edge and end devices. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, pp. 328–339. IEEE, 2017.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lampe, G. Llama: Open and efficient foundation language models, 2023a. URL <https://arxiv.org/abs/2302.13971>.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Yang, F., Bai, Y., Li, T., Wang, P., Liu, X., Dong, L., Zhang, Z., Zhou, G., et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- Yu, G.-I., Jeong, J. S., Kim, G.-W., Kim, S., and Chun, B.-G. Orca: A distributed serving system for transformer-based generative models. In *Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 521–538, 2022.
- Zeng, L., Chen, X., Zhou, Z., Yang, L., and Zhang, J. Co-edge: Cooperative dnn inference with adaptive workload partitioning over heterogeneous edge devices. *IEEE/ACM Transactions on Networking*, 29(2):595–608, 2020.
- Zhang, S. Q., Lin, J., and Zhang, Q. Adaptive distributed convolutional neural network inference at the network edge with adcnn. In *Proceedings of the 49th International Conference on Parallel Processing*, pp. 1–11, 2020.
- Zhang, S. Q., Lin, J., and Zhang, Q. A multi-agent reinforcement learning approach for efficient client selection in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 9091–9099, 2022.
- Zhang, S. Q., Lin, J., Zhang, Q., and Chen, Y.-J. Learning client selection strategy for federated learning across heterogeneous mobile devices. In *2024 25th International Symposium on Quality Electronic Design (ISQED)*, pp. 1–7. IEEE, 2024.
- Zhong, Y., Liu, S., Chen, J., Hu, J., Zhu, Y., Liu, X., Jin, X., and Zhang, H. Distserve: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2024.