

ConServe: Harvesting GPUs for Low-Latency and High-Throughput Large Language Model Serving

Yifan Qiao Shu Anzai Shan Yu Haoran Ma Yang Wang[‡]
 Miryung Kim Harry Xu
UCLA [‡]*Intel*

Abstract

Many applications are leveraging large language models (LLMs) for complex tasks, and they generally demand low inference latency and high serving throughput for interactive online jobs such as chatbots. However, the **tight latency requirement** and **high load variance of applications** pose challenges to serving systems in achieving high GPU utilization. Due to the high costs of scheduling and preemption, **today’s systems generally use separate clusters to serve online and offline inference tasks**, and dedicate GPUs for online inferences to avoid interference. This approach leads to **underutilized GPUs** because **one must reserve enough GPU resources for the peak expected load**, even if the average load is low.

This paper proposes to **harvest stranded GPU resources for offline LLM inference tasks** such as document summarization and LLM benchmarking. Unlike online inferences, these tasks usually **run in a batch-processing manner with loose latency requirements**, making them a **good fit for stranded resources that are only available shortly**. To enable safe and efficient GPU harvesting without interfering with online tasks, we built **ConServe**, an LLM serving system that contains (1) an **execution engine that preempts running offline tasks upon the arrival of online tasks**, (2) an **incremental checkpointing mechanism** that **minimizes the amount of recomputation required by preemptions**, and (3) a **scheduler that adaptively batches offline tasks** for higher GPU utilization. Our evaluation demonstrates that ConServe achieves strong performance isolation when co-serving online and offline tasks but at a much higher GPU utilization. When colocating practical online and offline workloads on popular models such as Llama-2-7B, ConServe achieves $2.35\times$ higher throughput than state-of-the-art online serving systems and reduces serving latency by $84\times$ compared to existing co-serving systems.

1 Introduction

Large language models (LLMs) such as ChatGPT [49], LLaMA [75], and GPT-4 [50] have emerged as transformative tools across a wide spectrum of application domains.

LLM-powered services, *e.g.*, interactive chatbot [9, 49, 73], programming assistant [18, 24, 65], and document summarization tools [39, 43], have already shown promises to businesses and end consumers [23] and are expected to create increasing impact in all aspects of human lives.

However, the superior abilities of LLMs come with high computational and memory demands to serve LLM inference requests. User requests to LLMs are served by running model inferences on GPUs, which can be significantly more expensive than typical web requests. To make matters worse, many LLM services such as chatbots are hosted *online*, requiring low response latency for user experience. To meet the strict latency service level objectives (SLOs), operators often need to overprovision GPUs to the LLM service, causing significant resource waste.

The tension between low tail latency and high resource utilization is further exaggerated by the **bursty load patterns** commonly exhibited in LLM workloads. LLM load varies not only over long timescales of hours, but also over timescales as short as a few seconds. For example, a recent study [78] reported that the load to ChatGPT can increase by $3\times$ within one minute. As a result, the service provider must provision GPUs to the peak user load to prevent vast latency SLO violations under load bursts.

A parallel trend is that **LLM-based applications are evolving into compound AI systems** [88], which involve retrieval-augmented generation (RAG) [58, 90], tool usage [31, 52], SQL-based data analytics [37], recommendation systems [77, 82], *etc.* Other than simply using LLMs to reply to online user requests, compound AI systems often **use LLMs for offline batch inference**, which has loose latency requirements but desires high generation throughput in a best-effort manner¹.

While many recent LLM serving systems are aimed at optimizing the inference efficiency [2, 22, 30, 57, 86, 91], they operate under the **assumption that the workload follows a constant request rate and latency requirement** and hence **may fall short when dealing with real-world workloads**, which are

¹For brevity, herein we refer to latency-critical requests as *online*, and best-effort requests that are latency insensitive as *offline*.

bursty and heterogeneous (*i.e.*, with both online and offline requests and drastically different SLOs). As a result, to deploy these systems, today’s datacenter operators must use separate clusters for serving online and offline requests, and overprovision GPUs for online serving to meet tight latency requirements [64].

The conventional wisdom to reduce resource waste is to dynamically repartition the cluster in response to the load variation of online requests [8, 33]. However, it requires *a priori* knowledge of the online load variation patterns, which are often hard to predict accurately. Allocation based on inaccurate load estimation can either be too conservative (still resulting in overprovisioning) or too radical (leading to SLO violations). In addition, re-allocating GPUs across different jobs is a complicated task that involves tearing down an existing process, cleaning up GPU resources, launching a new process with adjusted parameters, and setting up GPUs for serving. These operations are often time-consuming and take seconds to minutes to finish, preventing the system from rapidly reacting to load bursts.

Insights. The key question we ask in this paper is: instead of partitioning the cluster and serving online and offline requests separately, can we *co-serve them with the same set of GPUs*? In other words, our goal is no longer to force GPU allocation to fit unpredictable online load variation, but instead to adapt offline serving throughput to the available GPU resources. In doing so, we can maximize the GPU utilization while shifting the need for GPU reallocation to offline serving, which can tolerate relatively high response latency.

We developed *ConServe*, a unified LLM serving system that serves online and offline requests simultaneously and dynamically coordinates GPU resources between them. ConServe frees the service provider from the burden of manually allocating and adjusting GPUs for online serving. Instead, it achieves high GPU utilization by opportunistically scheduling offline requests whenever GPU compute resources and memory are available. To avoid resource contention that may break the latency SLO of online serving, ConServe proactively preempts offline requests and rapidly reclaims resources in response to online load bursts. ConServe recovers preempted requests after online requests are handled.

Challenges. The idea of co-locating latency-critical jobs with best-effort jobs has been widely studied in traditional cloud workload scheduling [8, 17, 55]. However, to realize its benefits in LLM serving, ConServe needs to overcome several unique challenges.

First, how can the system quickly free up resources taken by offline serving in response to online load bursts? Since LLM inference is iterative, a natural idea is to preempt offline requests after a generation iteration. However, the online request rate is highly unpredictable and they come with a tight latency requirement (usually in milliseconds). Serving a large batch of offline requests, even for a single iteration, may block incoming online requests for seconds.

To solve this problem, ConServe preempts running offline requests at the (fine) granularity of model layers. We found that the layer granularity strikes a good balance between responsiveness and execution efficiency. On the one hand, for different LLMs their layer sizes are generally small and do not vary much, allowing ConServe to discard partial results and reclaim occupied GPU resources much faster than finishing an entire iteration. On the other hand, compared to choosing a finer granularity such as CUDA kernels [21], instrumenting a model on a per-layer basis is lightweight, incurring only negligible overhead (see §4.3).

Second, how can the system minimize the recomputation cost? While the system can react to online load bursts with preemption, it discards the intermediate states (*i.e.*, KV cache) of the preempted offline requests. Consequently, it requires expensive recomputation to recover the lost KV cache once the preempted requests are resumed. A natural idea is to swap out the KV cache to host memory. Unfortunately, the size of the KV cache can grow unbounded, and swapping a large amount of data may, again, block incoming online requests.

Our insight is that LLM inference is stateless, and the KV cache remains unchanged once generated. Leveraging this property, ConServe incrementally checkpoints the KV cache of offline requests to the host memory per generation iteration. Since each request only generates one token per iteration, the amount of data to be written to the host memory is small and consistently bounded. Additionally, ConServe performs checkpointing and swap-in asynchronously, overlapping them with computation to improve efficiency (see §4.4).

Finally, how can the system preserve the latency SLO for running online requests? While batching offline requests improves GPU utilization, the batch size must be adjusted dynamically to ensure that online requests in the same batch can meet the latency SLO. This is, however, challenging because the inference latency is the result of many factors including the batch size, the total number of tokens, the phase of each request (*i.e.*, prefill or decode), *etc.* To overcome this challenge, ConServe uses a profiler and a SLO-aware scheduler. The profiler runs offline and collects the execution time of different input batch sizes and input lengths for requests in different stages. The scheduler adaptively changes the number of offline requests and tokens based upon the profiling results and the number of online tokens (see §4.5).

Results. We evaluated ConServe with the Llama-2-7B model [75] on two real-world datasets and synthetic workloads. Our results demonstrate that ConServe significantly improves GPU utilization while maintaining low online latency and high offline throughput. Compared to the state-of-the-art LLM serving system vLLM [30], which serves online and offline inferences separately, ConServe can achieve $2.35\times$ higher throughput with comparable online latency. Furthermore, when compared to existing co-serving solutions, ConServe provides strong performance isolation for online inference and yields an $84\times$ reduction in tail latency.

2 Background

2.1 Large Language Model Inference

Today’s LLMs typically generate outputs following the *autoregressive* process—they take a sequence of tokens as input, and repeatedly predict the next token given the input sequence and all the previously generated tokens. This process involves two distinct phases: `prefill` and `decode`. An LLM inference starts with the `prefill` phase that deals with a new input sequence and generates the first output token. Thanks to the modern model architecture such as Transformer [76], an LLM can process all input tokens in parallel, making the `prefill` phase compute-bound. After the `prefill` phase, the LLM enters the `decode` phase that sequentially generates subsequent tokens. Because each `decode` iteration generates only one token per request, it is less compute-intensive and typically bounded by the GPU memory bandwidth due to the need to load model weights. To improve its compute intensity and hardware efficiency, a common practice is to batch multiple requests in a single inference run to reuse the model weights for computation and memory load.

More specifically, during LLM inference, each token has its own intermediate state, represented by a series of key vectors and value vectors. As token states remain unchanged throughout the inference process, the inference engine [45, 47, 54] caches them in GPU memory (a.k.a., KV cache) to avoid recomputation in each `decode` step. However, due to the large size of KV vectors and the excessive number of tokens in a batch, the KV cache can consume a significant amount of GPU memory, which limits the batch size and hence hardware efficiency.

2.2 Characterizing LLM Serving

The rise of compound AI systems and LLM agents dictates that LLMs be used for various demands and in different forms. Generally, LLM serving can be categorized into two types: online serving, which generates responses to user inputs in real time, and offline serving, which processes user inputs and generates outputs in batches. Typical scenarios for offline serving including document summarization [27], data wrangling [42], LLM-enhanced data analytics [37], LLM benchmarking and evaluation [34], among other emerging applications. However, online and offline serving exhibit drastically different characteristics, as elaborated below.

Online serving is mostly suitable for latency-critical requests. Unlike traditional cloud services, where processing times are often predictable, a request to an LLM generates a sequence of tokens with varying lengths in multiple steps. As a result, LLM serving latency is measured on a per-token basis. Specifically, because LLM inference consists of two distinct phases, an LLM’s serving latency is defined by two key metrics: *time to first token (TTFT)*, which is the duration of the `prefill` phase, and *time per output token (TPOT)*,

which is the execution time of each step in the `decode` phase.² To provide a low end-to-end response latency and smooth user experience, it is critical for an online LLM serving system to optimize both TTFT and TPOT under stringent SLOs. For example, an online chatbot may set a 99th percentile TTFT SLO of 1.5 seconds and a 99th percentile TPOT SLO of 100 milliseconds to respond faster than a typical human can read [41, 60].

In contrast, offline serving is better suited for best-effort requests that are insensitive to response latency. Users typically submit offline requests in batch, while the serving system processes them in a best-effort manner to maximize hardware efficiency. Unlike online serving that emphasizes low latency, offline serving often involves processing large corpora or datasets, where the first-order performance metric is *throughput* that measures how many tokens are generated per second.

The different service-level performance objectives between online and offline serving poses challenges to LLM serving systems. To overcome the tension between low response latency (TTFT and TPOT) and high generation throughput, today’s LLM serving systems usually adopt different configurations and designs for different serving scenarios, typically requiring the service provider to set up separate clusters for online and offline serving. Next, we will discuss how existing systems optimize LLM serving and why they fall short in achieving high GPU utilization.

2.3 Existing LLM Serving Systems

A large body of system optimizations have been proposed to serve LLMs with low latency and high throughput since the launch of ChatGPT.

Throughput-Oriented Optimizations. Orca [86] is one of the pioneer LLM serving systems that employ *continuous batching* to improve inference throughput. It adds incoming requests directly into existing running batches, achieving higher throughput with a larger batch size. However, because new requests are in their `prefill` phase while running requests are in their `decode` phase, Orca sacrifices TPOT due to the larger batch size and additional `prefill` computation. In this same direction, vLLM [30] further increases the batch size by reducing the KV cache fragmentation and hence the GPU memory consumption.

Latency-Oriented Optimizations. Recently, more systems have been proposed to optimize LLM inference latency. Among them, Sarathi-Serve [2] and FastGen [22] piggyback on the continuous batching strategy, but break the `prefill` phase of incoming requests into small, fixed-size chunks (chunked-`prefill`), and only add one chunk into the running batch in each step. This limits the cost of additional `prefill`-phase com-

²In this paper, we measure TPOT as the generation time of each decode token (also referred to as inter-token latency in some literature) rather than the average token generation latency of each request.

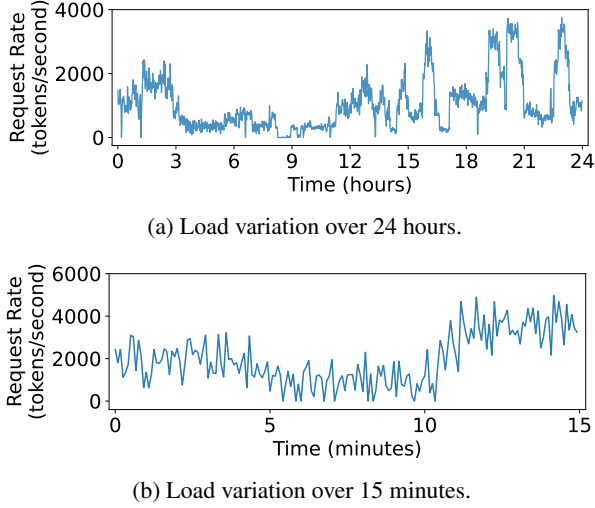


Figure 1: User traffic to ChatGPT within a campus exposes high load variability at various time scales.

putation in each generation step, thereby improving TPOT. However, as a new request is broken into many small chunks and processed in multiple steps, this line of work often leads to sacrificed TTFT.

Another line of work moves away from continuous batching scheduling, as exemplified by DistServe [91], Splitwise [57]. Instead, they duplicate the model and disaggregate the *prefill* and the *decode* computation onto different GPUs. Specifically, new requests are sent to a dedicated cluster for the *prefill* computation, and the generated token and KV cache are populated to another dedicated cluster that performs the *decode* computation. Llumnix [71] is another recent work that migrates KV caches between servers to reduce load imbalance. These systems achieve low TTFT and TPOT at the cost of more GPUs. Due to the additional data transfer of KV caches between servers, they also lead to reduced throughput.

Limitations of Existing Techniques. Existing LLM serving systems suffer from two major limitations. First, they trade off between latency and throughput, and hence are only applicable for either online or offline serving. Second, they can neither determine how many GPUs should be provisioned, nor adaptively change the number of GPUs during serving. In fact, because they target only one type of incoming requests, they often assume a constant request arrival rate and reserve enough GPUs in advance.

Unfortunately, these problems are more pronounced when serving real-world inference workloads, which can contain both online and offline requests with high load variation. Such a challenge makes it hard to provision the right amount of GPU resources *a priori*.

3 Motivation

In this section, we first demonstrate how real-world LLM load can vary over time and why existing serving systems fall short

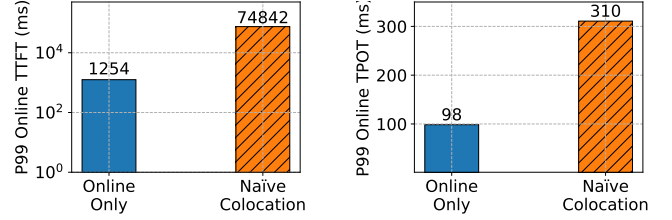


Figure 2: 99th-percentile TTFT and TPOT of online requests when co-served with offline requests using a priority-based scheduler. Note that the y-axis in the left figure is on a log scale.

of achieving high GPU utilization. We then use an experiment to quantitatively demonstrate why simply co-locating online serving with offline serving cannot solve the problem.

Online Load Burstiness. Real-world LLM workloads often expose diurnal patterns and high load variability, as backed by a recent study [78] which collects user traffic to ChatGPT for two months within a campus. Figure 1a shows the load variation within a day. Despite the average load being as low as 1050 tokens per second, there is a clear contrast between peak hours and non-peak hours. The load can achieve more than 3743 tokens per second in the afternoon, while in the morning the service only experiences little traffic. In addition, there are unpredictable load bursts within an hour, during which the request rate can ramp up multiple times in a short period. Figure 1b further provides a closer examination of one such burst in a 15-minute window. As shown, the load still fluctuates drastically at the minute scale, and the request rate increases by $3\times$ in the tenth minute.

Due to such high load variability, service operators must reserve resources for peak demand to avoid violating latency SLOs. However, since the peak inference load can be multiple times higher than the average load, overprovisioning can lead to significant resource waste.

Naïve colocation hurts online serving latency. A simple strawman approach is to extend today’s online serving systems to allow users to submit requests with priorities. Users may assign online requests a high priority and offline requests a low priority. The scheduler should incorporate the priority information and schedule online requests first to preserve low latency. Since none of the existing systems support co-serving online and offline requests, we implemented a priority-based scheduler atop vLLM [30] (details in §6.1) and used it to serve Llama-2-7B on one NVIDIA A100 GPU. We replayed the trace in Figure 1 and collected the 99th-percentile TTFT and TPOT for online requests. These results are reported in Figure 2.

While co-serving offline and online requests has indeed improved GPU utilization, it could significantly impact online serving latency—P99 TTFT increases by $59.7\times$ and P99 TPOT increases by $3.16\times$. There are two reasons for the latency increase. First, once offline requests are scheduled, they are batched together with online requests and cannot be pre-empted selectively. Therefore, incoming online requests must

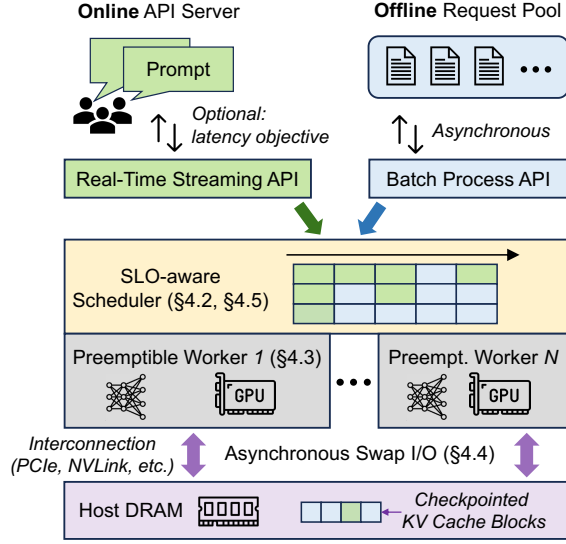


Figure 3: Design overview of ConServe. ConServe efficiently co-serves online and offline requests with three major components: an SLO-aware scheduler, a set of preemptive workers, and an incremental checkpointing mechanism.

wait until they are served, leading to significantly increased queuing delays. Second, the scheduler tends to pack enough offline requests to make full use of GPU memory, which can lead to large batch sizes that take longer to finish, and hence a further increased delay.

These problems call for a new system that can harvest available GPU resources for offline serving *on-the-fly* and dynamically adjust the allocation of resources to different types of requests efficiently and safely. To achieve efficiency, the system should be able to swiftly scale offline serving up and down with minimal disruption. To ensure safety, the system must respond quickly to load bursts and preserve online serving SLOs.

4 Design

4.1 ConServe Overview

ConServe is an LLM serving system designed to co-serve online and offline requests. Figure 3 shows its overall architecture. At its frontend, ConServe provides similar APIs to other LLM serving systems. For online requests, it uses a real-time streaming API that returns outputs once each token is generated. For offline requests, it adopts an interface similar to OpenAI’s Batch API [51], which takes a batch of requests from a pool and returns responses asynchronously.

ConServe relies on its backend to schedule and execute inference jobs, which consists of three major components. The scheduler runs as a daemon thread and continuously fetches and schedules offline requests. Upon the arrival of an online request, the scheduler reactively preempts offline requests and immediately schedules the incoming online request (§4.2).

The scheduler leverages a set of workers to host the LLM and serve the scheduled batch. A model can span multiple GPUs, where each GPU is managed by one worker. ConServe supports both tensor and pipeline parallelism for high efficiency and flexibility (§4.3). To quickly recover preempted requests with minimized recomputation costs, ConServe leverages host memory and incrementally checkpoints KV caches during its execution (§4.4). Finally, because ConServe can schedule both online and offline requests in the same batch, it adopts an SLO-aware policy to adaptively adjust the batch size to preserve online TTFT and TPOT latency objectives while maximizing the hardware efficiency (§4.5).

4.2 Unified Preemptive Scheduler

ConServe’s scheduler is the key component that serves both online and offline requests in a unified fashion. As with previous LLM serving schedulers [30, 86], ConServe adopts continuous batching in its scheduler. To prevent long prefills from interfering decode iterations, ConServe also adopts chunked prefill [2] that partitions and computes long prefills into small chunks over multiple iterations and limits the batch size per iteration.

The overall scheduling logic is shown in Algorithm 1. It maintains two separate queues for online and offline requests, and continuously monitors the online queue for incoming requests. In each scheduling step, it first calculates a budget batch size given the latency objectives for TTFT and TPOT (Line 10). The budget limits both the number of tokens and the number of requests in a batch. The detailed policies are elaborated in §4.5. The scheduler then checks and schedules incoming online requests within the budget allowance (Lines 11-15). To prevent previously scheduled offline requests from blocking incoming online requests (*i.e.*, head-of-line blocking), the scheduler excludes offline tokens from the budget first (Line 12-13), and then reactively preempts offline requests if the budget is over-saturated (Line 16). This preemption continues until all scheduled requests can fit into the budget (Lines 27-34). After accommodating all online requests, the scheduler opportunistically schedules offline requests using the remaining budget (Line 17-19). The detailed scheduling policies (*i.e.*, SLOAWARESCHEDULE) to maintain online latency SLOs and manage GPU memory usage will be discussed shortly in §4.5.

Offline Batching Mode. Due to the diurnal load patterns, a model may not receive any new online requests periodically during non-peak hours. During such periods, the scheduler switches to the *offline batching mode* for maximizing offline serving throughput (Line 20-22 in Algorithm 1). Because offline requests come with only loose or no latency requirements, the scheduler ignores the budget limit and sets the largest batch size that can saturate GPU compute or memory capacity.

Algorithm 1: The unified scheduling logic to co-serve online and offline requests.

Input: time to first token (TTFT) objective t_{TTFT} , time per output token (TPOT) objective t_{TPOT} .

```

1 Global  $Q_{on}$  (online request queue)
2 Global  $Q_{off}$  (offline request queue)
3 Global  $Q_{out}$  (output queue)
4 Global  $t_{sched}$  (time when last batch gets scheduled)
5 Function UNIFIEDSCHEDULE ( $Q_{on}$ ):
6   Initialize token budget  $\tau \leftarrow Inf$ 
7   Initialize current scheduled batch  $B \leftarrow \emptyset$ 
8   Initialize new online requests has_new_online  $\leftarrow$  False
9   while True do
10     $\tau \leftarrow \text{calc\_budget}(t_{TTFT}, t_{TPOT}) - B.\text{num\_tokens}()$ 
11    if  $!Q_{on}.\text{is\_empty}()$  then
12       $\tau_{off} \leftarrow B.\text{num\_offline\_tokens}()$ 
13       $B_{on}, \tau \leftarrow \text{SLOAWARESCHEDULE}(Q_{on}, \tau + \tau_{off})$ 
14       $B \leftarrow B \cup B_{on}$ 
15      has_new_online  $\leftarrow$  True
16     $B, \tau \leftarrow \text{PREEMPTOVERBUDGETOFFLINE}(B, \tau)$ 
17    if  $B.\text{has\_online\_requests}()$  then
18       $B_{off}, \tau \leftarrow \text{SLOAWARESCHEDULE}(Q_{off}, \tau)$ 
19       $B \leftarrow B \cup B_{off}$ 
20    else
21       $B_{off}, \tau' \leftarrow \text{SLOAWARESCHEDULE}(Q_{off}, Inf)$ 
22       $B \leftarrow B \cup B_{off}$ 
23     $t_{sched} \leftarrow \text{time.now}()$ 
24     $B, B_{finished} \leftarrow \text{exec\_batch}(B)$ 
25     $Q_{out}.\text{append}(B_{finished})$ 
26 Function PREEMPTOVERBUDGETOFFLINE ( $B, \tau$ ):
27   if  $\tau.\text{over\_budget}()$  then
28     for  $R$  in  $B.\text{offline\_reqs}()$  do
29       PREEMPTSCHEDULING ( $R$ )
30        $B \leftarrow B \setminus \{R\}$ 
31        $\tau \leftarrow \tau - R.\text{num\_tokens}()$ 
32       if not  $\tau.\text{over\_budget}()$  then
33         break
34   return  $B, \tau$ 

```

Preemption. ConServe can preempt offline requests at two potential points: during scheduling or model execution. First, in each scheduling step, the scheduler needs to preempt scheduled offline requests to make room for incoming online requests or free up GPU memory under memory pressure. Similar to prior systems such as vLLM [30], preempting a request during scheduling can be done by either discarding-and-recomputing or swapping to host memory, and it is implemented in the PREEMPTSCHEDULING function at Line 29 in Algorithm 1.

However, in the worst case where incoming online requests are about to exceed their TTFT objectives, ConServe must

Algorithm 2: Preemptive scheduling logic: the arrival of an online request triggers a callback function and preempts running workers if necessary to meet the TTFT objective.

Input: Incoming online request o .

```

1 Global  $Q_{on}$  (online request queue)
2 Global  $t_{sched}$  (time when last batch gets scheduled)
3 Function ONRECVONLINEREQUEST ( $o$ ):
4   Initialize current time  $t_{curr} \leftarrow \text{time.now}()$ 
5    $Q_{on}.\text{append}(o)$ 
6    $B \leftarrow \text{Worker.get\_curr\_batch}()$ 
7    $t_{exec} \leftarrow \text{Profiler.estimate\_exec\_time}(Q_{on} \cup B)$ 
8    $t_{est} \leftarrow \text{Profiler.estimate\_exec\_time}(B)$ 
9    $t_{remain} \leftarrow t_{est} - (t_{curr} - t_{sched})$ 
10  if  $t_{remain} + t_{exec} > t_{TTFT}$  then
11    break
12   $B_{victim} \leftarrow \emptyset$ 
13  for  $R$  in  $B.\text{offline\_reqs}()$  do
14     $B \leftarrow B \setminus \{R\}$ 
15     $B_{victim} \leftarrow B_{victim} \cup \{R\}$ 
16     $t_{exec} \leftarrow \text{Profiler.estimate\_exec\_time}(Q_{on} \cup B)$ 
17     $t_{est} \leftarrow \text{Profiler.estimate\_exec\_time}(B)$ 
18     $t_{remain} \leftarrow t_{est} - (t_{curr} - t_{sched})$ 
19    if  $t_{remain} + t_{exec} \leq t_{TTFT}$  then
20      break
21  PREEMPTRUNNING ( $B_{victim}$ )

```

deal with the urgent case and preempt offline requests even if they are in a running batch to schedule new requests in a timely fashion. To this end, ConServe scheduler invokes an asynchronous handler upon the arrival of new online requests, as shown in Algorithm 2. The handler first pushes the incoming online request into the online request queue, and then leverages the profiler (discussed shortly in §4.5) to estimate the queuing delay and the execution time (Lines 7-9). If the estimated serving time exceeds the TTFT objective, the scheduler signals the worker to preempt offline requests in the current running batch until it can meet the TTFT objective (Lines 10-21).

4.3 Preemptible Worker

To host the model on GPUs and execute inference requests, ConServe employs a set of workers to manage GPU resources. Each worker is responsible for handling the physical KV cache, launching model computation, and communicating with other workers. ConServe supports both tensor and pipeline parallelism in the Megatron-LM fashion [70]. Additionally, ConServe’s workers support preempting a running batch (*i.e.*, PREEMPTRUNNING in Algorithm 2 Line 21) as the last resort to avoid head-of-line blocking. This is particularly crucial in offline batching mode, where large batch sizes are used for maximized throughput at the cost of high latency.

However, preempting model execution while it is running involves several challenges in balancing responsiveness and runtime overhead.

The first challenge is how the scheduler should notify all workers. When the model is deployed with tensor parallelism, workers perform collective communications. Preempting a subset of workers indiscriminately may leave others waiting during communication operations, potentially causing the program to hang. To prevent this, ConServe periodically synchronizes all workers using a distributed barrier [62] and signals preemption only after synchronization. The ConServe scheduler shares a preemption flag with its workers, which is set when it decides to preempt the current batch, and checked by workers during their execution. Upon detecting the flag is set. Workers check this flag during execution, and upon detecting it, they immediately abort the remaining execution and return an empty result. The scheduler then cleans up the partial computation states, restores its own state, and quickly schedules any incoming online requests.

The second challenge is determining the appropriate temporal granularity for preemption. If the granularity is too coarse, the system may not respond timely to load bursts; conversely, if it is too fine-grained (*e.g.*, per GPU kernel as in previous work [21]), synchronization overhead may become significant and reduce overall efficiency. To strike a balance, ConServe chooses to preempt at the granularity of *model layers* for two reasons. First, while architectures and sizes of model may vary drastically, LLMs all consist of multiple layers. The execution time for a single layer accounts for only a small fraction of the end-to-end latency, which allows for responsive preemption. Second, LLM layers involve dozens of GPU kernels and communication operations, so the overhead of preemption and synchronization is small relative to the layer execution time.

To avoid impacting online serving latency, ConServe restrict layer-wise preemption to the offline batching mode. In co-serving mode, where the batch contains both online and offline requests, ConServe leverages chunked prefill and SLO-aware scheduling (discussed shortly) to limit chunk sizes and bound execution time per iteration. While in offline batching mode, ConServe aggressively increases batch sizes and chunk sizes and relies on layer-wise preemption to maintain responsiveness.

To further mitigates costs for small models whose layer executes fast, ConServe allows workers to adjust the preemption granularity by batching multiple layers before reaching a synchronization barrier. In our evaluation, preempting per eight layers achieves a favorable balance of low runtime cost and high responsiveness (details in §6.4.2).

4.4 Asynchronous Swap I/O

While preempting offline requests can free up GPU compute resources immediately, ConServe must also address the GPU

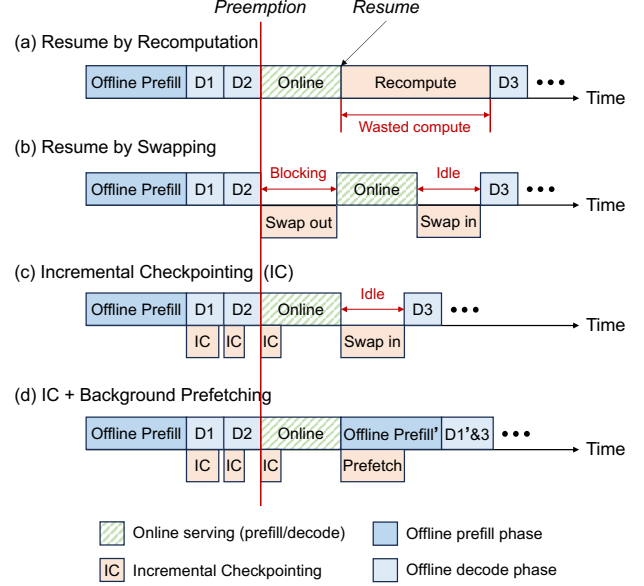


Figure 4: Comparison between different resume strategies. (a) Resume by recomputation achieves low preemption delay at the cost of additional computation. (b) Resume by swapping reduces the recomputation cost but swapping out can block the schedule of incoming online requests. (c) Incremental checkpointing (IC) minimizes both preemption delay and resume cost. (d) IC + background swap-in overlaps swap-in with prefill computation of the next batch and achieves consistently high GPU utilization.

memory pressure caused by the excessive KV cache usage of offline serving. To quickly reclaim memory in response during online load bursts and minimize the scheduling delay, one strawman approach is to directly discard KV caches for victim offline requests that are preempted, and recompute them later once online load subsides (see Figure 4(a)). However, this approach sacrifices GPU efficiency by wasting compute resources on redundant recomputation, thereby reducing overall serving throughput.

To avoid expensive recomputation, a common optimization applied by existing serving systems such as vLLM [30] is to swap out KV caches of victim requests to host memory, as shown in Figure 4(b). However, swapping out by itself still takes time, which delays the scheduling of incoming online requests. To make matters worse, the amount of data to be swapped out increases proportionally to the number of tokens in offline requests, while the interconnection bandwidth between GPUs and host memory is limited. For example, an Nvidia A100 GPU connecting to host DRAM via PCIe 4.0x16 offers only 32 GBps of bandwidth. Swapping out all KV caches for large offline requests can easily take dozens or even thousands of milliseconds and significantly block online requests. Moreover, when resuming preempted offline requests, the GPU must swap in the evicted KV cache blocks before continuing the computation, leaving the GPU idle and again, hurting overall serving throughput.

Incremental Checkpointing. To overcome these inefficiencies, we propose a novel asynchronous checkpointing and resuming mechanism. Firstly, instead of swapping KV caches out at the last minute when preemption happens, ConServe worker adopts a checkpointing mechanism that runs asynchronously and incrementally in the background. As shown in Figure 4(c), ConServe incrementally checkpoints newly generated KV caches of offline requests to host memory after each generation iteration. This is enabled by the autoregressive nature of modern LLMs, which generate tokens iteratively. As a result, the device-host I/O traffic is amortized over multiple iterations. Besides, because checkpointing has no data dependency with follow-up computation, it can be done asynchronously in the background and overlapped with computation. As a result, it incurs only negligible runtime overhead.

ConServe manages GPU memory similar to vLLM [30] by reserving GPU memory ahead of time and virtually mapping KV cache blocks to physical GPU memory. But unlike swapping which frees KV caches right after they are swapped out, checkpointing keeps KV caches in GPU memory until incoming online requests are scheduled. Discarding KV caches in ConServe is as fast and lightweight as freeing victim KV cache blocks and remapping new ones virtually, which finishes in dozens of microseconds.

Background Prefetching. Secondly, because offline requests do not have strict latency constraints, they offer ConServe workers a unique opportunity to re-order requests and overlap swap-in with computation. As shown in Figure 4(d), instead of waiting for victim requests to be swapped in, ConServe worker launches a new offline batch and runs for its prefill phase, and meanwhile prefetches KV cache blocks in the background. Afterward, the worker will merge the new batch with prefetched requests and run their decode phase together. For long sequences that need to swap in a large amount of KV blocks to resume, ConServe will also split the swap-in phase over multiple steps to avoid long swap-in delays. In this way, ConServe keeps device-host I/O for offline requests entirely in the background and eliminates idle GPU cycles.

Adaptive Checkpointing Policy. While checkpointing and prefetching can greatly improve GPU efficiency, blindly applying them without coordination will excessively use host memory and interconnection bandwidth, potentially causing resource contention when the computation is lightweight or online requests also need swapping. Furthermore, checkpointing is only necessary under GPU resource pressure when offline requests are likely to be preempted. In other cases, it can be avoided to save host memory and reduce runtime overhead.

Inspired by the asynchronous swap system design in the OS kernel [63] and the conventional wisdom of random early detection [15], ConServe adaptively controls the checkpointing rate based on GPU memory pressure to limit resources used by checkpointing, and it gradually increases the checkpoint-

ing rate as GPU memory usage goes up and always tries to match checkpointing speed with memory consumption speed. Specifically, ConServe starts checkpointing when available GPU memory is running low (by default the threshold is set to 50%). but it will only checkpoint a small number of offline requests first and gradually increase the number of offline requests to checkpoint after observing constantly increasing memory usage. This is sufficient for most cases when GPU memory pressure is moderate, and ConServe can selectively preempt offline requests that have been checkpointed. Under the extreme case when memory pressure still persists after preempting all checkpointed requests, ConServe will discard the KV cache for the remaining offline requests to ensure incoming online requests can get scheduled without any delay.

Finally, incremental checkpointing can also help online serving to accommodate swapping/preemption caused by continuous batching. With the continuous batching policy, the scheduler will eagerly add new requests to the batch to enlarge the batch size when GPU memory permits, but this may consume all GPU memory when requests in the batch generate long output sequences and keep allocating memory for KV caches. In such scenarios, the scheduler will have to swap or preempt online requests to make room. ConServe will also incrementally checkpoint online requests under GPU memory pressure, thereby eliminating the cost of swapping out or recomputation for online requests as well.

4.5 SLO-aware Scheduling

The last challenge that ConServe must address is to determine the right *batch size* (i.e., the number of offline requests and tokens) to compute and the right *set of KV cache blocks* to checkpoint/prefetch for each inference iteration. (1) As for the batch size, if the scheduler batches too few offline requests, a substantial portion of GPU compute resources and memory bandwidth would be underutilized and lead to suboptimal overall serving throughput; on the flip side, if the scheduler batches too many offline requests, they will saturate the GPU and increase the inference computation time and hence online serving latencies, resulting in SLO violations. To make matters more complicated, the model execution time depends on not only the batch size but also the context lengths of each request due to the non-linear compute complexity of the attention kernel ($O(n^2)$ for prefill phase and $O(n)$ for decode phase where n is the context length)³. Therefore, the scheduler must also decide the number of offline tokens that can be batched and processed together with online tokens while meeting the SLO. (2) As for the KV cache blocks to checkpoint/prefetch, swapping too many blocks, even if in the background, will exhaust the host-device bandwidth and GPU streaming multiprocessors (SMs) resources and block

³Chunked prefill alleviates this effect by splitting long sequences into smaller chunks, but its performance is still prone to prefix lengths due to repeated KV cache access [2].

the computation kernel, while checkpointing too few blocks will make it unable to keep up with the GPU memory consumption rate, leading to memory exhaustion and triggering swap that blocks incoming online requests.

To tackle this challenge, ConServe adopts an SLO-aware scheduler to collect the model execution profiles with an *offline profiler* and dynamically adjust the batch size and the degree of background swapping leveraging the collected information and SLOs specified by the users. To flexibly batch varying numbers of offline tokens into each batch, ConServe leverages chunked prefill [2] to break offline sequences if necessary.

Profiler. ConServe’s performance and concrete scheduling policy are highly dependent on the model itself and the hardware configurations, including specific GPU models, the parallelization strategy, PCIe bandwidth, *etc.* To quantify the performance impact, ConServe will first run its profiler in the offline phase to profile the model prefill and decode computation time with different numbers of tokens, as well as the swap latency with respect to the number of KV cache blocks. The profiled results will be saved locally and automatically loaded when launching a ConServe server.

SLO-aware Policy. ConServe then leverages the profiled information to schedule offline requests. For each scheduled online batch, it queries the profiler with the latency SLO (TPOT for batches containing decode phase requests, TTFT otherwise) to get the maximum number of tokens that can be processed. It then schedules just enough offline tokens to fulfill the batch. Similarly, it uses the SLO to decide the maximum number of KV cache blocks that can be swapped in the background, and defers the extra blocks to the next round. In many cases where swapping a block is faster than computation, or offline tokens only take a small portion of the batch size, this policy is memory-safe and can always checkpoint KV cache blocks faster than the GPU memory consumption. Under the extreme case where a huge amount of KV cache blocks need to be checkpointed, ConServe will prioritize online latency SLO attainment and discard excessive KV cache blocks and recompute them later.

5 Implementation

We have implemented ConServe atop vLLM 0.4.2 [30] with 4165 lines of code. ConServe also leverages Ray [40] to distribute the model to multiple GPUs, but it additionally supports a Python `multiprocessing` backend that communicates via shared memory to eliminate Ray’s high inter-process communication (IPC) cost when running with multiple GPUs on a single node.

To unify the scheduling logic, ConServe implements its online and offline request queues as priority queues with two priority levels so they can share the same scheduler code but follow strict priority orders. ConServe’s real-time streaming API will set requests to a high priority, while the batch

API sets requests to low priority internally, and users are not required to manually specify priorities for requests.

To support layer-wise preemption, ConServe instruments the model to add a preemption *safe point* between layers, inspired by the design of managed language runtimes [53]. The safe point contains a small piece of code that synchronizes all workers and then checks the `preempted` flag, which is a variable shared by the scheduler and the master worker. Upon preemption happens, the master worker will broadcast the `preempted` flag to the other workers using NCCL [44]. To ensure layer-wise preemption is only enabled in offline batching mode, ConServe’s scheduler passes workers an additional `preemptible` flag to activate all safe points after scheduling a pure offline batch. Otherwise, safe points are disabled and incur zero runtime overhead.

To support asynchronous checkpointing, ConServe enhances vLLM’s KV cache manager by keeping track of the mapping between each GPU KV block and its corresponding CPU KV block, which store the checkpoint. This mapping is recorded in an extended field of the virtual page table, allowing it to be queried and updated by the ConServe scheduler.

ConServe supports flexible and customizable adaptive checkpointing policies through two key interfaces provided by its KV cache manager: `checkpoint(seqs: List[Sequence])` and `get_blocks_to_chkpt()->List[KVBlock]`. After each execution step, the scheduler calls `checkpoint()` for sequences that were previously executed, marking them as checkpointing candidates. However, the actual decision to checkpoint is deferred until before scheduling the next batch of requests, when the scheduler invokes `get_blocks_to_chkpt()`. At this point, the KV cache manager applies the policy to select all or a subset of candidate KV blocks for checkpointing. This design not only allows users to customize the policy according to their needs but also allows ConServe to skip checkpointing when GPU memory is sufficient.

To ensure overlap between computation and asynchronous swap I/O, ConServe runs model computation on the default CUDA stream while assigning a separate CUDA stream for KV cache checkpointing and prefetching in each worker. For asynchronous swap I/Os, ConServe launches swap kernels in their dedicated stream before executing each model layer, thus pipelining and overlapping entire swap I/O with computation. Since checkpointed or prefetched KV cache blocks have no data dependencies with the ongoing model computation, no synchronization is required between the two CUDA streams.

ConServe also comes with a built-in load generator that can generate precisely timed requests following the gamma distribution. The load generator can be configured with various parameters including the request rate, burstiness (*i.e.*, skewness of the gamma distribution), and request lengths to cover the characteristics of real-world workload.

6 Evaluation

Our evaluation aims to answer the following questions:

1. Can ConServe judiciously coordinate GPU resources between online and offline serving to optimize overall performance? (§6.2)
2. Can ConServe quickly and reactively harvest available idle GPU resources to improve offline serving throughput? (§6.3.1)
3. Can ConServe quickly and reactively react to online load bursts and maintain low latency? (§6.3.1 and §6.3.2)
4. What contributes to ConServe’s better performance? (§6.4)

6.1 Setup

Environment and Model. We conducted experiments on one server that equips a 48-core CPU, 320GB GB memory, and one NVIDIA A100-40G GPU. The server ran Ubuntu 22.04 and CUDA 12.1. To reduce latency jitter, we disabled dynamic voltage and frequency scaling (DVFS) of GPUs [46, 72] and Python garbage collector [61].

We evaluated ConServe on Llama-2-7B model [75]. All experiments use FP16 precision and tensor parallelism to align with our baselines.

Baselines. We compared ConServe with vLLM [30], a state-of-the-art LLM serving system. We also enabled chunked-prefill in vLLM to ensure a fair comparison. Because the original vLLM cannot co-serve online requests and offline requests, we evaluated it by only feeding online requests (referred to as *Online-Only*), which provides the optimal online serving latency but zero offline serving throughput. To enhance its performance, we also extended vLLM’s online serving frontend with a batch process API, so that it can also take batched offline requests while serving online requests. We additionally implemented a priority scheduler that prioritizes online requests over offline ones, and we refer to this enhanced baseline as *vLLM++*.

Real Workloads. To model the bursty load patterns of online requests, we evaluated ConServe and the other baseline systems with a real-world load trace BurstGPT [78], which collects user requests to ChatGPT [49] and GPT-4 [50] in a university campus and is representative for online workloads. To adapt the campus-wide trace to our evaluation setting that consists of a relatively small number of GPUs, we sample the trace following the previous practice [68] as follows: given the duration D and request rate R , we sample $R \times D$ requests from the original trace, and re-scale the real-time stamps to $[0, D]$. We set R as the maximal request rate that can be served within latency SLOs.

For offline workloads, we evaluated the document summarization task with the LongBench [6] datasets.

Synthetic Workloads. To demonstrate ConServe’s capability in reacting to different request rates and different degrees

of load burstiness and SLO tightness, we also leverage ConServe’s load generator to generate synthetic workloads with configurable parameters (details in §6.3).

Metrics. Since online serving targets low latency and offline serving targets high throughput, we adopt different metrics when evaluating their service quality. For online serving, we measure each request’s 99th percentile TTFT and TPOT, respectively. For offline serving, we measure the throughput by counting the number of generated tokens per second.

6.2 Overall Serving Performance

In this section, we evaluated ConServe with real-world workloads and compared its end-to-end performance against the baselines. We used the same trace reported in Figure 1b as the online load. We first ran the original vLLM with only online loads to collect its 99th percentile TTFT and TPOT, and then set them as the SLO targets for all systems (1500ms for TTFT and 110ms for TPOT).

A good result for ConServe would show that it quickly detects any GPU underutilization and judiciously batches offline requests to harvest available GPU resources to achieve good offline throughput while still keeping online serving latency lower than the SLO. In contrast, *Online-Only* should achieve the optimal online serving latency but fails to harvest idle GPU resources for offline serving. On the contrary, *vLLM++* does batch offline requests together with online ones, but it optimizes for the overall serving throughput and does not guarantee online latency. Therefore, we expect *vLLM++* to achieve high offline throughput but also experience drastically fluctuating TTFT and TPOT during load bursts.

Figure 5 presents the results. The top two figures present 99th percentile TTFT and TPOT of all three systems, with red horizontal lines indicating the SLOs for TTFT (1500ms) and TPOT (110ms), respectively. The bottom figure shows the overall (online + offline) serving throughput. As expected, *Online-Only* achieves lowest TTFT and TPOT. However, during periods of low request rate (from $t = 240$ s to 620s), it cannot batch enough requests per inference iteration, resulting in low latency but underutilizing the GPU. On average, it achieves an overall serving throughput of 1999 tokens/s. ConServe maintains low online TTFT and TPOT that are close to optimal ones and consistently lower than SLOs, and it offers 3702 tokens/s overall throughput by harvesting available GPU resources for offline serving. While *vLLM++* is also able to batch offline requests and achieves the highest throughput at 4308 tokens/s, frequent swapping significantly increases its 99th percentile TTFT and TPOT by $84 \times$ (83825ms) and $25 \times$ (2523ms), respectively.

In summary, the experiment demonstrates that ConServe can efficiently utilize available GPU resources for offline serving with negligible impact on online serving latency. As a result, it achieves $2.35 \times$ higher overall throughput compared

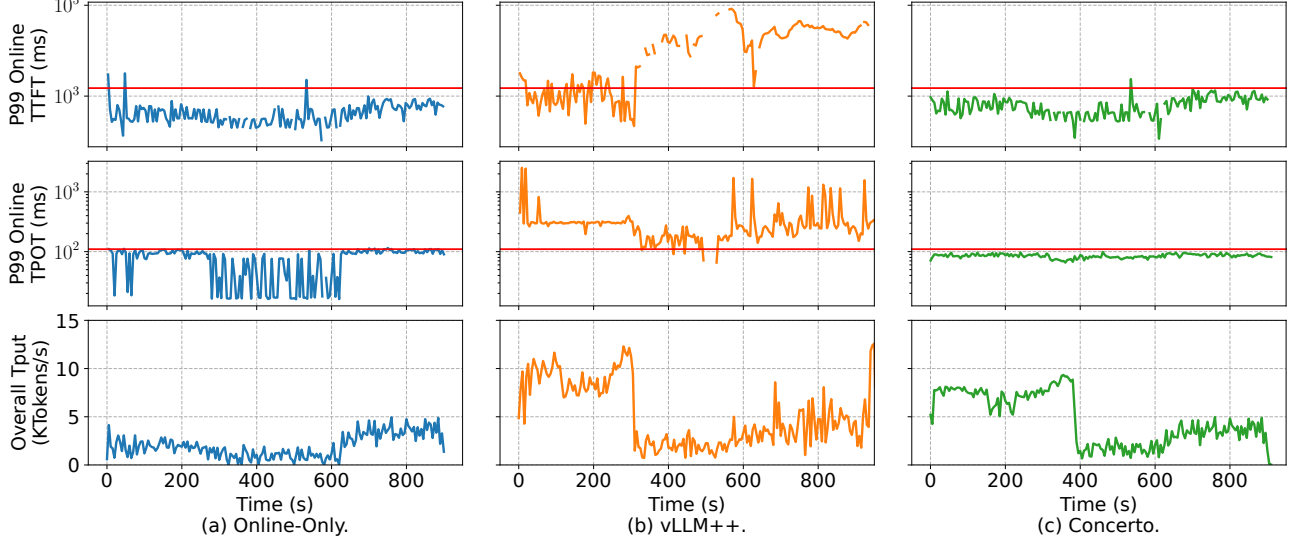


Figure 5: Overall serving performance on real workloads. ConSerto achieves consistently low TTFT and TPOT that are comparable with Online-Only and below the SLO. It also achieves 86% of the ideal offline serving throughput (measured by vLLM++ which eagerly batches offline requests regardless online latency constraints).

to Online-Only and 98.8% lower online serving latency compared to vLLM++.

6.3 Reacting to Load Bursts

In this section, we conducted a set of experiments to investigate whether ConSerto can reactively harvest idle GPU resources—whenever they are available—for offline serving, while still quickly reacting to resource pressure to avoid interfering with online serving performance. For experiments in this section, we set the request input length to 1024 and the output length to 128 as representative values for online loads [78, 91].

6.3.1 ON/OFF Phased Load Patterns

In many real-world settings, LLMs do not always receive requests from clients. Instead, they may remain idle for a while (“OFF” phases), and experience high load occasionally (“ON” phases) [68]. To evaluate whether ConSerto can react quickly even under intense resource pressure caused by online load bursts, we synthesized a staged online load by dynamically changing the load between the system’s max capacity and zero, as shown as the blue line in Figure 6. A good result for ConSerto will show that it can quickly react to changes in resource availability and keep both low online tail latency and high GPU utilization.

Figure 6 presents the results. Initially, the system ran in the ON stage at its maximum capacity. At $t = 180$ s, the system switched to the OFF stage with no online load until $t = 360$ s, when it returned to the ON stage. Managing such resource pressure is particularly challenging, as load spikes can occur instantly. Despite these sharp changes between ON and OFF

stages, ConSerto is still able to avoid SLO violations and keep the 99th percentile TTFT and TPOT under 350ms and 90ms, respectively. Additionally, ConSerto quickly and reactively detects idle GPU resources and shifts to offline serving at millisecond-scale, achieving an offline throughput of 5868 tokens/s during OFF stages. vLLM++, in contrast, batches offline requests too aggressively during ON stages, resulting in $1.4\times$ to $11\times$ higher 99th percentile TTFT and TPOT along with vast SLO violations.

These results show that ConSerto can effectively harvest idle GPU resources for offline serving and keep GPUs at maximum utilization. More importantly, ConSerto can quickly and reactively preempt offline requests under pressure. Its memory reclamation speed exceeds the rate at which online serving can allocate. Consequently, ConSerto can always safely harvest GPU resources while the online load neither runs out of resources nor experiences slowdowns.

6.3.2 Robustness to Changing Load Burstiness

In reality, clients’ behavior often follows dynamic and unpredictable load patterns [68, 78]. To this end, we further investigated the robustness of ConSerto under varying degrees of load burstiness and request rates. Following previous studies [33, 78], we constructed a synthetic load following the Gamma process with an average rate of 2 requests per second and a coefficient of variation (CV) of 1. We used CV to measure the load burstiness, where larger CV values indicates greater load variability. We examined how the online 99th percentile latencies are affected when either CV or request rate is held constant while the other is varied. As modeled by queueing theory [14], we expect the queueing delay to

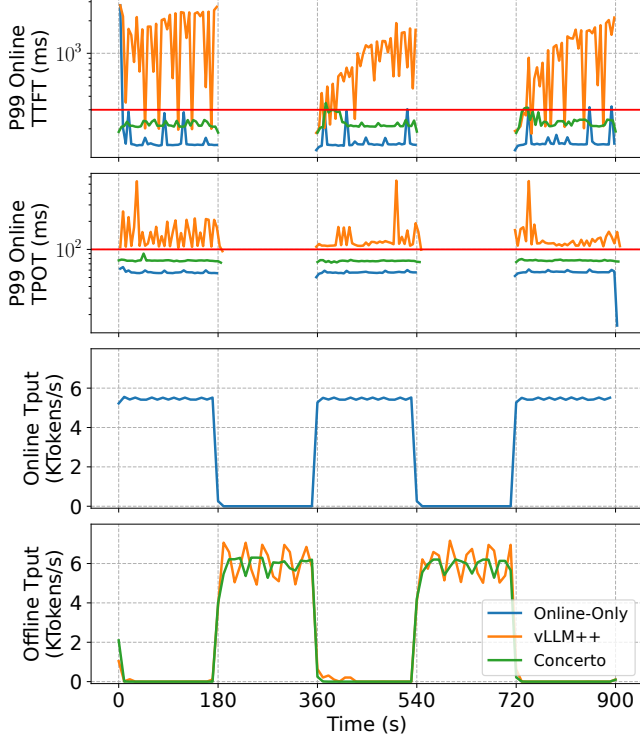


Figure 6: Overall serving performance on ON/OFF phased workloads. ConServe incurs negligible impact on online TTFT and TPOT and always keeps them below their SLO during ON phases. During the transition from the ON to OFF stages, ConServe reactively detects and harvests additional idle resources and achieves high offline throughput during the OFF stages. Even under extreme resource pressure during the transition from the OFF to ON stages, ConServe quickly scales down offline serving and prevents any spikes in online latency.

increase superlinearly as either CV or request rate increases. However, ConServe should be able to handle uncertain resource pressure and maintain tail latencies comparable to the ideal ones, even in highly bursty scenarios.

Figure 7 presents 99th percentile online TTFTs, 99th percentile online TPOTs, and offline throughputs achieved by all three systems under varying CVs (left column) and varying request rates (right column). As expected, online TTFT increases when the load becomes more bursty and heavier across all systems. However, ConServe demonstrates strong resilience to bursty loads and high request rates. It maintains low TTFTs that are close (within 25%) to the ideal latencies achieved by Online-Only. In contrast, vLLM++ suffers from significantly higher online TTFTs, with a minimum of 4980ms, making it difficult to meet latency SLOs. Surprisingly, despite batching fewer requests per iteration than vLLM++ to keep latencies low, ConServe can still outperform vLLM++ by 4%–12% in terms of offline throughput. A detailed examination reveals that vLLM++ suffers from frequent I/O stalls caused by swapping and underutilizes GPU com-

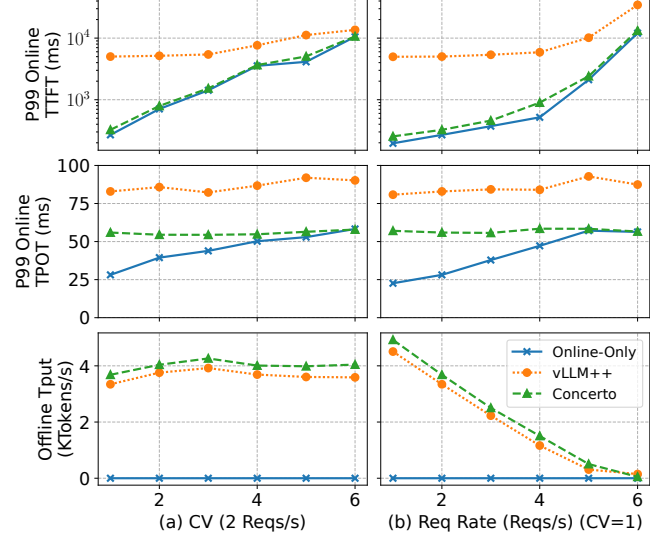


Figure 7: Overall serving performance under varying CVs and request rates. ConServe consistently achieves low online latencies and enables a linear trade-off between online throughput and offline throughput to keep maximized GPU utilization.

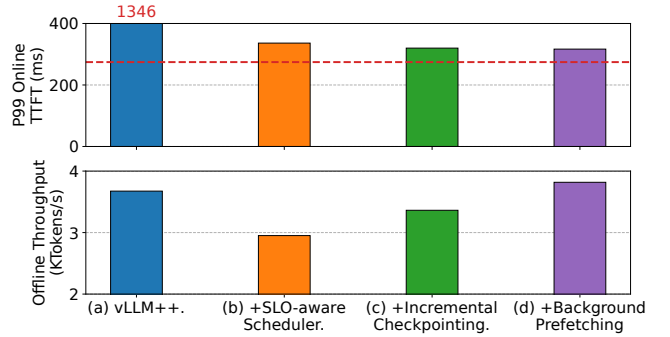


Figure 8: All ConServe’s optimizations work in tandem to improve performance.

pute, whereas ConServe eliminates these stall by overlapping checkpointing and swap-ins for offline requests.

In summary, these results demonstrate that ConServe can consistently handle intense resource pressure and avoid on-line latency SLO violations. It maintains robust performance across a wide range of request rates and varying degrees of load burstiness.

6.4 Design Drill-Down

We now evaluate specific aspects of ConServe’s design to understand their individual contributions to overall performance.

6.4.1 Breaking Down End-to-End Speedup

we re-ran the experiment under the same conditions as Figure 7 with a fixed CV of 1 and a request rate of 2 request/s. We incrementally enabled optimizations and reported results

in Figure 8. We also included `Online-Only` results (the red dashed line) as a reference. As shown in the leftmost bar, the naïve `vLLM++` baseline achieves an offline throughput of 3674 tokens/s but increases the 99th percentile TTFT to 1346ms. By contrast, ConServe’s preemptive and SLO-aware scheduler reduces the 99th TTFT by 71.4% to 446ms, bringing it within 26% of the ideal online latency. However, this optimization also reduces the offline throughput to 2951 tokens/s for two reasons: (1) preemptive scheduling triggers frequent swapping events; and (2) the SLO-aware policy limits batch size when online and offline requests co-exist to preserve online latency at the cost of lower offline throughput. Enabling incremental checkpointing and background prefetching recovers 14.0% and 13.6% of the throughput loss, respectively, ultimately achieving an offline throughput of 3818 tokens/s. With all optimizations enabled, ConServe is able to reduce 99th percentile online TTFT by 76.5% while improving the offline throughput by 1.04 \times .

6.4.2 Efficiency of Preemptible Worker

We measured the runtime cost of the preemption safepoint and compared it to the model execution time. On average, each instrumented safepoint takes 988 μ s. Our profiling reveals that most of this overhead is due to PyTorch’s distributed barrier, which we plan to optimize by implementing a custom lightweight barrier. In comparison, the model execution takes 98.5ms on average.

Nevertheless, ConServe only enables safepoints for pure offline batches, and online serving will never be affected. To strike a balance between the runtime overhead and the responsiveness, ConServe instruments the model every 8 layers. When no preemption occurs, this instrumentation introduces a total latency increase of 3.99ms, which accounts for 4% of the model’s execution time. Meanwhile, ConServe maintains responsiveness by detecting new requests and preempting the running batch within 5.41ms, ensuring timely responses to incoming online requests.

7 Discussion

Compatibility with Disaggregated LLM Serving. In latency-sensitive scenarios, users may have stringent TTFT and TPOT requirements. Many ongoing research efforts have proposed disaggregated architectures for LLM serving [57, 91] that use separate GPUs for prefill computation and decode computation to reduce interference. ConServe’s design is orthogonal and compatible with these disaggregated architectures. Specifically, ConServe can be integrated separately in the prefill cluster scheduler and the decode cluster scheduler in a disaggregated architecture. Furthermore, ConServe’s offline serving interfaces expose more semantics to reduce the KV cache transfer cost between prefill clusters and

decode clusters. For example, ConServe can not only checkpoint to local DRAM but also asynchronously checkpoint offline KV cache from the prefill cluster to the GPU memory or DRAM in decode clusters. We leave these optimizations as future work.

Long-Context Scenarios. Improving LLMs’ ability to process long contexts and long outputs has recently gained significant attraction [32, 36, 83]. ConServe is compatible with sequence parallelism. Optimizations such as RingAttention [36] and StreamLLM [83] remain valid, in which case ConServe will partition both online and offline requests among all sequence parallel workers for load balancing.

Support for Multiple Models. While the main design goal of ConServe is to co-serve online and offline requests *within* a model to reuse the model weights, it can also be generalized to serve multiple models. For example, in practice, many models can share model weights but adapt to different tasks with parameter-efficient fine-tuning (PEFT) [7, 67, 81]. ConServe can seamlessly support them and flexibly co-serve online and offline requests for different fine-tuned models.

8 Related Work

Model Serving Systems. Other than specialized LLM serving systems discussed in §2.3, many other systems target serving more general ML models. Triton [48], TorchServe [56, 74], TensorFlow Serving [1, 19] are three representative serving systems ready for production. Clipper [11], InferLine [10], and Clockwork [20] serve general neural networks by batching and scheduling requests. REEF [21] and SHEPHERD [89] proposed to co-locate models on the same GPU and preempt GPU compute kernels for scheduling. AlpaServe [33], on the other hand, leverages model parallelism for statistical multiplexing. However, these systems overlook the huge model size and the autoregressive nature of LLM inference, hence only achieving suboptimal LLM serving performance.

Offline LLM Serving. As offline inference has gained increasing traction, many systems are specifically optimized for offline LLM serving. DeepSpeed ZeRO-Inference [4] and FlexGen [69] proposed to offload model weights and KV caches to host memory to serve LLMs on small commodity GPUs. S³ [28] optimized for high generation throughput by predicting the output sequence length and minimizing memory waste. Their design does not fit online serving due to the long swapping latency, but they can benefit from ConServe’s incremental checkpointing and background prefetching mechanism for further performance improvement.

Optimized LLM Algorithms and Architectures. Another line of work focuses on optimizing the efficiency of LLM algorithms/kernels. FlashAttention [12, 13] improves the memory I/O efficiency with a redesigned attention kernel. GPTQ [16], AWQ [35], and SqueezeLLM [29] quantize and compress the model weights and KV caches to reduce GPU memory

consumption. Some other work aims to improve compute and memory efficiency with optimized transformer architectures. MQA [66] and GQA [3] modify the attention kernel to reduce the KV cache size. Mixture-of-expert models [5, 26, 38] make weight parameters sparse and hence reduce the model size. ConServe is orthogonal to these optimizations on algorithms and architectures, while ConServe can further improve GPU utilization beyond their benefits.

Deep Learning Schedulers. GPU clusters today suffer from low resource utilization [25, 79, 85]. To improve GPU utilization, many deep learning schedulers [79, 80, 84, 85, 87] are proposed for better model placement, job migration, and GPU sharing. ConServe focuses on sharing GPU resources between online and offline serving, and it is orthogonal to cluster-level schedulers. However, ConServe could benefit from advances in underlying scheduling policies and hardware support.

Workload Co-location. Co-locating latency-critical applications with batch applications is a widely adopted approach to improve resource utilization in datacenters. For instance, Parties [8] partitions resources such as CPU cache and memory resources across microservices to preserve their SLOs. Operating systems such as ZygOS [59], Shenango [55], and Caladan [17] proposed to preempt batch jobs and reallocate CPU cores to latency-critical jobs to improve CPU utilization. However, existing workload co-location solutions primarily target traditional hardware resources and workloads, rather than emerging GPUs and AI workloads. In contrast, ConServe shares the concept of colocating workloads and preemption to improve resource utilization but is specifically optimized for LLM serving on GPUs.

9 Conclusion

We present ConServe, a unified LLM serving system that serves both online and offline inference and maintains high GPU utilization, low latency, and high throughput simultaneously. ConServe achieves low latency through its unified scheduler, which opportunistically schedules offline requests when GPU resources are available, and preemptive worker, which timely preempts offline requests to mitigate interference before it can harm online latency. ConServe further incrementally and asynchronously checkpoints KV caches for offline requests to minimize the recomputation cost, and adopts an SLO-aware scheduler to maximize throughput.

References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [2] A. Agrawal, N. Kedia, A. Panwar, J. Mohan, N. Kwatra, B. S. Gulavani, A. Tumanov, and R. Ramjee. Taming throughput-latency tradeoff in llm inference with sarathi-serve, 2024.
- [3] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023.
- [4] R. Y. Aminabadi, S. Rajbhandari, M. Zhang, A. A. Awan, C. Li, D. Li, E. Zheng, J. Rasley, S. Smith, O. Ruwase, and Y. He. Deepspeed inference: Enabling efficient inference of transformer models at unprecedented scale, 2022.
- [5] M. Artetxe, S. Bhosale, N. Goyal, T. Mihaylov, M. Ott, S. Shleifer, X. V. Lin, J. Du, S. Iyer, R. Pasunuru, G. Anantharaman, X. Li, S. Chen, H. Akin, M. Baines, L. Martin, X. Zhou, P. S. Koura, B. O’Horo, J. Wang, L. Zettlemoyer, M. Diab, Z. Kozareva, and V. Stoyanov. Efficient large scale language modeling with mixtures of experts, 2022.
- [6] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, Y. Dong, J. Tang, and J. Li. Longbench: A bilingual, multitask benchmark for long context understanding, 2024.
- [7] L. Chen, Z. Ye, Y. Wu, D. Zhuo, L. Ceze, and A. Krishnamurthy. Punica: Multi-tenant lora serving, 2023.
- [8] S. Chen, C. Delimitrou, and J. F. Martínez. Parties: Qos-aware resource partitioning for multiple interactive services. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS ’19*, page 107–120, New York, NY, USA, 2019. Association for Computing Machinery.
- [9] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.
- [10] D. Crankshaw, G.-E. Sela, X. Mo, C. Zumar, I. Stoica, J. Gonzalez, and A. Tumanov. Inferline: latency-aware provisioning and scaling for prediction serving pipelines. In *Proceedings of the 11th ACM Symposium on Cloud Computing, SoCC ’20*, page 477–491, New York, NY, USA, 2020. Association for Computing Machinery.
- [11] D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gonzalez, and I. Stoica. Clipper: A Low-Latency online prediction serving system. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 613–627, Boston, MA, Mar. 2017. USENIX Association.

- [12] T. Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.
- [13] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [14] C. Delimitrou and C. Kozyrakis. Amdahl’s law for tail latency. *Commun. ACM*, 61(8):65–72, jul 2018.
- [15] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, 1993.
- [16] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023.
- [17] J. Fried, Z. Ruan, A. Ousterhout, and A. Belay. Caladan: Mitigating interference at microsecond timescales. In *OSDI*. USENIX Association, 2020.
- [18] GitHub. Github copilot - write code faster. <https://copilot.github.com/>, 2021.
- [19] Google. Tensorflow serving is a flexible, high-performance serving system for machine learning models. <https://www.tensorflow.org/tfx/guide/serving>, 2024.
- [20] A. Gujarati, R. Karimi, S. Alzayat, W. Hao, A. Kaufmann, Y. Vigfusson, and J. Mace. Serving DNNs like clockwork: Performance predictability from the bottom up. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pages 443–462. USENIX Association, Nov. 2020.
- [21] M. Han, H. Zhang, R. Chen, and H. Chen. Microsecond-scale preemption for concurrent GPU-accelerated DNN inferences. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 539–558, Carlsbad, CA, July 2022. USENIX Association.
- [22] C. Holmes, M. Tanaka, M. Wyatt, A. A. Awan, J. Rasley, S. Rajbhandari, R. Y. Aminabadi, H. Qin, A. Bakhtiari, L. Kurilenko, and Y. He. DeepSpeed-fastgen: High-throughput text generation for llms via mii and deepspeed-inference, 2024.
- [23] K. Hu. Chatgpt sets record for fastest-growing user base. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>, 2023.
- [24] N. Jain, T. Zhang, W.-L. Chiang, J. E. Gonzalez, K. Sen, and I. Stoica. Llm-assisted code cleaning for training accurate code generators, 2023.
- [25] M. Jeon, S. Venkataraman, A. Phanishayee, J. Qian, W. Xiao, and F. Yang. Analysis of Large-Scale Multi-Tenant GPU clusters for DNN training workloads. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, pages 947–960, Renton, WA, July 2019. USENIX Association.
- [26] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts, 2024.
- [27] H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods, 2024.
- [28] Y. Jin, C.-F. Wu, D. Brooks, and G.-Y. Wei. S3: increasing gpu utilization during generative inference for higher throughput. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2024. Curran Associates Inc.
- [29] S. Kim, C. Hooper, A. Gholami, Z. Dong, X. Li, S. Shen, M. W. Mahoney, and K. Keutzer. Squeezellm: Dense-and-sparse quantization, 2024.
- [30] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP ’23*, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery.
- [31] LangChain. Langchain: Build context-aware reasoning application. <https://python.langchain.com/>, 2024.
- [32] D. Li, R. Shao, A. Xie, E. P. Xing, X. Ma, I. Stoica, J. E. Gonzalez, and H. Zhang. Distflashattn: Distributed memory-efficient attention for long-context llms training, 2024.
- [33] Z. Li, L. Zheng, Y. Zhong, V. Liu, Y. Sheng, X. Jin, Y. Huang, Z. Chen, H. Zhang, J. E. Gonzalez, and I. Stoica. AlpaServe: Statistical multiplexing with model parallelism for deep learning serving. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 663–679, Boston, MA, July 2023. USENIX Association.

- [34] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. Holistic evaluation of language models, 2023.
- [35] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han. Awq: Activation-aware weight quantization for llm compression and acceleration, 2024.
- [36] H. Liu, M. Zaharia, and P. Abbeel. Ring attention with blockwise transformers for near-infinite context, 2023.
- [37] S. Liu, A. Biswal, A. Cheng, X. Mo, S. Cao, J. E. Gonzalez, I. Stoica, and M. Zaharia. Optimizing llm queries in relational workloads, 2024.
- [38] Mistral AI. Mixtral-8x22b: Cheaper, better, faster, stronger. <https://mistral.ai/news/mixtral-8x22b/>, 2024.
- [39] Moonshot AI. AI Assistant with Memory. <https://www.perplexity.ai/>, 2024.
- [40] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, and I. Stoica. Ray: A distributed framework for emerging AI applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 561–577, Carlsbad, CA, Oct. 2018. USENIX Association.
- [41] Mosaic AI Research. Llm inference performance engineering: Best practices. <https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices>, 2024.
- [42] A. Narayan, I. Chami, L. Orr, S. Arora, and C. Ré. Can foundation models wrangle your data?, 2022.
- [43] S. Narayan, S. B. Cohen, and M. Lapata. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.
- [44] NVIDIA. Collective Communications Library (NCCL). <https://developer.nvidia.com/nccl>, 2024.
- [45] NVIDIA. Fastertransformer: Transformer related optimization, including bert, gpt. <https://github.com/NVIDIA/FasterTransformer>, 2024.
- [46] NVIDIA. System Management Interface SMI. <https://developer.download.nvidia.com/compute/DCGM/docs/nvidia-smi-367.38.pdf>, 2024.
- [47] NVIDIA. Tensorrt-llm: A tensorrt toolbox for optimized large language model inference. <https://github.com/NVIDIA/TensorRT-LLM>, 2024.
- [48] NVIDIA. Triton Inference Server. <https://developer.nvidia.com/triton-inference-server>, 2024.
- [49] OpenAI. Chatgpt: Conversational language model. <https://chat.openai.com>, 2023.
- [50] OpenAI. Gpt-4 technical report, 2023.
- [51] OpenAI. Batch api. <https://platform.openai.com/docs/guides/batch/batch-api>, 2024.
- [52] OpenAI. Introducing the gpt store. <https://openai.com/index/introducing-the-gpt-store/>, 2024.
- [53] OpenJDK. HotSpot Glossary of Terms: Safe-point. <https://openjdk.org/groups/hotspot/docs/HotSpotGlossary.html>, 2024.
- [54] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling, 2019.
- [55] A. Ousterhout, J. Fried, J. Behrens, A. Belay, and H. Balakrishnan. Shenango: Achieving high CPU efficiency for latency-sensitive datacenter workloads. In *NSDI*, pages 361–378, 2019.
- [56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.
- [57] P. Patel, E. Choukse, C. Zhang, A. Shah, Íñigo Goiri, S. Maleki, and R. Bianchini. Splitwise: Efficient generative llm inference using phase splitting, 2024.
- [58] Perplexity AI. Perplexity is a free ai search engine. <https://www.perplexity.ai/>, 2024.
- [59] G. Prekas, M. Kogias, and E. Bugnion. Zygos: Achieving low tail latency for microsecond-scale networked tasks. In *Proceedings of the 26th Symposium on Operating Systems Principles, SOSP ’17*, page 325–341, New York, NY, USA, 2017. Association for Computing Machinery.

- [60] Proxet. Llm has a performance problem inherent to its architecture: Latency. <https://www.proxet.com/blog/llm-has-a-performance-problem-inherent-to-its-architecture-latency>, 2023.
- [61] Python Software Foundation. gc—Garbage Collector interface. <https://docs.python.org/3/library/gc.html>, 2024.
- [62] PyTorch. Distributed communication package: Barrier. <https://pytorch.org/docs/stable/distributed.html#torch.distributed.barrier>, 2024.
- [63] Y. Qiao, C. Wang, Z. Ruan, A. Belay, Q. Lu, Y. Zhang, M. Kim, and G. H. Xu. Hermit: Low-Latency, High-Throughput, and transparent remote memory via Feedback-Directed asynchrony. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 181–198, Boston, MA, Apr. 2023. USENIX Association.
- [64] R. Qin, Z. Li, W. He, M. Zhang, Y. Wu, W. Zheng, and X. Xu. Mooncake: A kvcache-centric disaggregated architecture for llm serving, 2024.
- [65] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Re-meze, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bit-ton, M. Bhatt, C. C. Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve. Code llama: Open foundation models for code, 2024.
- [66] N. Shazeer. Fast transformer decoding: One write-head is all you need, 2019.
- [67] Y. Sheng, S. Cao, D. Li, C. Hooper, N. Lee, S. Yang, C. Chou, B. Zhu, L. Zheng, K. Keutzer, J. E. Gonzalez, and I. Stoica. S-lora: Serving thousands of concurrent lora adapters, 2024.
- [68] Y. Sheng, S. Cao, D. Li, B. Zhu, Z. Li, D. Zhuo, J. E. Gonzalez, and I. Stoica. Fairness in serving large language models. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 965–988, Santa Clara, CA, July 2024. USENIX Association.
- [69] Y. Sheng, L. Zheng, B. Yuan, Z. Li, M. Ryabinin, B. Chen, P. Liang, C. Ré, I. Stoica, and C. Zhang. Flex-gen: high-throughput generative inference of large language models with a single gpu. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- [70] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro. Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- [71] B. Sun, Z. Huang, H. Zhao, W. Xiao, X. Zhang, Y. Li, and W. Lin. Llumnix: Dynamic scheduling for large language model serving, 2024.
- [72] Z. Tang, Y. Wang, Q. Wang, and X. Chu. The impact of gpu dvfs on the energy and performance of deep learning: an empirical study. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems, e-Energy ’19*, page 315–325, New York, NY, USA, 2019. Association for Computing Machinery.
- [73] The Claude Team. Introducing the next generation of claude. <https://www.anthropic.com/news/claude-3-family>, 2024.
- [74] The PyTorch Foundation. Torchserve is a performant, flexible, and easy to use tool for serving pytorch models in production. <https://pytorch.org/serve/>, 2024.
- [75] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.
- [76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.
- [77] J. Wang, H. Lu, Y. Liu, H. Ma, Y. Wang, Y. Gu, S. Zhang, S. Bi, L. Baugher, E. Chi, et al. Llms for user interest exploration in large-scale recommendation systems. *arXiv preprint arXiv:2405.16363*, 2024.
- [78] Y. Wang, Y. Chen, Z. Li, Z. Tang, R. Guo, X. Wang, Q. Wang, A. C. Zhou, and X. Chu. Towards efficient and reliable llm serving: A real-world workload study, 2024.
- [79] Q. Weng, L. Yang, Y. Yu, W. Wang, X. Tang, G. Yang, and L. Zhang. Beware of fragmentation: Scheduling GPU-Sharing workloads with fragmentation gradient descent. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, pages 995–1008, Boston, MA, July 2023. USENIX Association.
- [80] B. Wu, Z. Zhang, Z. Bai, X. Liu, and X. Jin. Transparent GPU sharing in container clouds for deep learning workloads. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 69–85, Boston, MA, Apr. 2023. USENIX Association.

- [81] B. Wu, R. Zhu, Z. Zhang, P. Sun, X. Liu, and X. Jin. dLoRA: Dynamically orchestrating requests and adapters for LoRA LLM serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 911–927, Santa Clara, CA, July 2024. USENIX Association.
- [82] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.
- [83] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks, 2024.
- [84] W. Xiao, R. Bhardwaj, R. Ramjee, M. Sivathanu, N. Kwatra, Z. Han, P. Patel, X. Peng, H. Zhao, Q. Zhang, F. Yang, and L. Zhou. Gandiva: Introspective cluster scheduling for deep learning. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 595–610, Carlsbad, CA, Oct. 2018. USENIX Association.
- [85] W. Xiao, S. Ren, Y. Li, Y. Zhang, P. Hou, Z. Li, Y. Feng, W. Lin, and Y. Jia. AntMan: Dynamic scaling on GPU clusters for deep learning. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pages 533–548. USENIX Association, Nov. 2020.
- [86] G.-I. Yu, J. S. Jeong, G.-W. Kim, S. Kim, and B.-G. Chun. Orca: A distributed serving system for Transformer-Based generative models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 521–538, Carlsbad, CA, July 2022. USENIX Association.
- [87] P. Yu and M. Chowdhury. Salus: Fine-grained gpu sharing primitives for deep learning applications, 2019.
- [88] M. Zaharia, O. Khattab, L. Chen, J. Q. Davis, H. Miller, C. Potts, J. Zou, M. Carbin, J. Frankle, N. Rao, and A. Ghodsi. The shift from models to compound ai systems. <https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/>, 2024.
- [89] H. Zhang, Y. Tang, A. Khandelwal, and I. Stoica. SHEPHERD: Serving DNNs in the wild. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 787–808, Boston, MA, Apr. 2023. USENIX Association.
- [90] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez. Raft: Adapting language model to domain specific rag, 2024.
- [91] Y. Zhong, S. Liu, J. Chen, J. Hu, Y. Zhu, X. Liu, X. Jin, and H. Zhang. Distserve: Disaggregating prefill and decoding for goodput-optimized large language model serving, 2024.