

---

# The Synergy of Speculative Decoding and Batching in Serving Large Language Models

---

Qidong Su<sup>123</sup> Christina Giannoula<sup>1</sup> Gennady Pekhimenko<sup>123</sup>

## Abstract

Large Language Models (LLMs) like GPT are state-of-the-art text generation models that provide significant assistance in daily routines. However, LLM execution is inherently sequential, since they only produce one token at a time, thus incurring low hardware utilization on modern GPUs. Batching and speculative decoding are two techniques to improve the GPU hardware utilization in LLM inference. To study their synergy, we implement a prototype implementation and perform an extensive characterization analysis on various LLM models and GPU architectures. We observe that the optimal speculation length depends on the batch size used. We analyze the key observation and build a quantitative model to explain it. Based on our analysis, we propose a new adaptive speculative decoding strategy that chooses the optimal speculation length for different batch sizes. Our evaluations show that our proposed method can achieve equal or better performance than the state-of-the-art speculation decoding schemes with fixed speculation length.

## 1. Introduction

Large Language Models (LLMs) such as GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), and LLama (Touvron et al., 2023) have achieved great success in not only natural language processing, but also other scientific and technical fields, including robotics (Brohan et al., 2023), programming languages (Chen et al., 2021), and medical science (Thirunavukarasu et al., 2023).

However, text generation, the most common usage of LLMs, is carried out in a sequential execution paradigm called auto-regression. In auto-regression, LLMs predict the subsequent *one* token based on previous tokens. The newly generated token will be appended to the end of the sequence of generated tokens. In each iteration, the input query tensor only

contains the embedding of one token (except for the first iteration), and only one token will be generated. This leads to a low amount of execution parallelism and GPU hardware utilization. To make things worse, the execution of LLM inference is usually memory-bounded (Ivanov et al., 2021). The sequential execution paradigm requires GPUs to load the huge weight matrices from off-chip memory to on-chip memory in each iteration, which leads to poor memory reuse and performance.

Researchers have explored two approaches to tackle the aforementioned problem: (i) batching (Yu et al., 2022; Fang et al., 2021), and (ii) speculative decoding (Chen et al., 2023; Leviathan et al., 2023; Miao et al., 2023; Spector & Re, 2023; Liu et al., 2023). First, batching, as used in other deep-learning applications (Gao et al., 2018; Crankshaw et al., 2017; Lee et al., 2018; Olston et al., 2017; Shen et al., 2019), merges several user requests together as one batched request, thus GPUs produce one new token in parallel for each user request in one batch. Second, speculative decoding uses a smaller model (Small Speculative Model, SSM) to predict more than one subsequent token (henceforth referred to as *speculation length*) and uses the original LLM to verify the prediction in parallel. Verifying a sequence of tokens is much faster than sequentially generating them thanks to the higher GPU hardware utilization. Although the SSM also provides subsequent tokens in a sequential manner, its overhead is much smaller than the runtime of LLM, thus it is amortized. Therefore, if the SSM can accurately mimic the behavior of the original LLM, the end-to-end runtime of inference can be significantly reduced.

Our goal in this work is to study the synergy of these two approaches and how we can get the best of these two approaches. Intuitively, they are both designed to saturate the GPU hardware resource, enabling one of them might offset the benefits of the other one. Enabling both of them simultaneously might be ineffective or even worse. Also, batching is not always available as it is constrained by the request traffic volume and available GPU memory size.

To this end, we provide an extensive characterization study of batching and speculative decoding execution. First, we carried out a comprehensive profiling study of batched speculative decoding using different models, GPUs, batch sizes,

---

<sup>1</sup>University of Toronto

<sup>2</sup>CentML Inc

<sup>3</sup>Vector Institute

and speculation lengths.

Our key observation is that the optimal speculation length varies across different batch sizes. More specifically, larger batch sizes require a smaller speculation length to achieve optimal performance, and a speculation length too large will deteriorate the performance. We developed an analytical model to explain this phenomenon. This necessitates the design of an adaptive strategy that finds the optimal configuration for different execution scenarios.

Therefore, we propose an adaptive speculative decoding strategy, which adjusts the speculation length according to the batch size used. It runs a short period of profiling before deployment and builds a mapping from the batch size to its corresponding optimal speculation length. After the service is launched, it will use the optimal speculation length for an arriving batch of requests.

Our evaluations show that our proposed method achieves equal or better performance than the state-of-the-art speculation decoding schemes with fixed speculation length. For time-varying requests, adaptive speculative decoding can provide an extra 9% latency reduction compared with fixed speculation length.

In conclusion, our major contributions are:

- We carry out a comprehensive study to investigate how the batch size interacts with the speculation length. We observe that the optimal speculation length depends on the batch size: smaller speculation lengths need to be chosen for large batch sizes.
- We analyze the observations of our study by building a quantitative model to explain them.
- We propose a new adaptive speculative decoding strategy that chooses the optimal speculation length for different batch sizes at low latency costs.
- We evaluate the proposed strategy and compare it with prior schemes with fixed speculation length. Our experiments show that our proposed method can achieve equal or better performance than the prior schemes.

## 2. Background

**Batching.** Batching (Yu et al., 2022) is one of the most popular performance optimizations in deep learning. It merges multiple data instances into one, to increase parallelism and hardware utilization. It can increase the throughput, especially when the batch size is small. In the context of LLMs, multiple user requests that arrive at roughly the same time can be processed in a batch. However, batching also demands more memory usage since the intermediate result of all requests in a batch needs to be stored. The effectiveness of batching is also bounded by the request traffic.

**Speculative Decoding.** Speculative decoding (Chen et al.,

2023; Leviathan et al., 2023; Miao et al., 2023) is another method to increase execution parallelism. It uses a Small Speculative Model (SSM) to first predict several subsequent tokens, and feed them into the original Large Language Model (LLM), which verifies the predictions and corrects incorrect predictions. The verification for each token can be parallelized, thus increasing the hardware utilization. The process of speculation decoding is shown in Algorithm 1.

In each step, the SSM predicts  $s$  subsequent tokens  $(t_1, t_2, \dots, t_s)$ , which will be verified by the LLM later. Verification is done by calculating the logits corresponding to each token  $(o_1, o_2, \dots, o_s)$ . The  $i$ -th logit  $o_i$  is a probability distribution across the whole vocabulary, representing which token should be the  $(i + 1)$ -th token, if all previous tokens are correct. If for some index  $l$ ,  $t_{l+1}$  is not the token with the highest probability in  $o_i$ , all tokens after  $t_l$  will be discarded, i.e. the correctness of one speculated token relies on the correctness of its previous tokens.

Assuming  $l$  of them are correct, the length of the correct sequence increases by  $l + 1$ , since the LLM provides either a correction or an extra look-ahead token, which is inferred from the logits of the last correct token. One point worth noting is that even if the SSM never produces a correct prediction, the algorithm can still terminate since the LLM can always generate one new correct token in each iteration ( $\text{argmin}(o_l)$  in the algorithm above).

---

### Algorithm 1 Speculative Decoding (argmax sampling)

---

**Input:** Input tokens  $I$ , speculation length  $s$   
 $S = I$   
Initialize KV cache  $kv_{SSM}, kv_{LLM}$   
**while true do**  
 $t_1, t_2, \dots, t_s = \text{speculate}(S, kv_{SSM}, s)$   
 $o_1, o_2, \dots, o_s = \text{LLM}(S, t_1, t_2, \dots, t_s, kv_{LLM})$   
 $l = \text{first index that } t_l \neq \text{argmin}(o_l)$   
 $S = S \oplus t_1, t_2, \dots, t_{l-1} \oplus \text{argmin}(o_l)$   
**if**  $\langle \text{EOS} \rangle \in S$  **then**  
    **break**  
**end if**  
**end while**  
**return**  $S$

---

## 3. Batched Speculative Decoding

In this section, we describe our implementation of batched speculative decoding and show the benchmark result of it for different models, GPUs, batch sizes, and speculation lengths. We also include empirical and modeling analysis to explain this phenomenon.

### 3.1. Prototype Implementation.

To investigate the synergy of speculative decoding and batching, we implemented a prototype implementation supporting both these optimizations for HuggingFace transformers models (Wolf et al., 2019). The correction of wrong speculation is implemented using attention masks: if one token does not pass the verification of the LLM, the current token and all subsequent tokens after it will be discarded by masking them off. The correct token given by the LLM is then appended to the end of the sequence, which will be used in the next iteration.

### 3.2. Experimental Analysis

We measure the average latency per token achieved for different combinations of batch sizes (from 1 to 32 with steps to be a power of 2) and speculation lengths (from 1 to 8) in various settings: (i) in Figure 1a, 1a, 1b we use RTX 3090 GPU model and vary the LLM model (OPT-1.3B, OPT-6.7B (Zhang et al., 2022) and Llama-7B (Touvron et al., 2023)), (ii) in Figure 1b, 1d, and 1c we use the OPT-6.7B model and we vary the GPU architecture (RTX 3090, RTX 4090, A100). In these figures, the asterisk annotated point corresponds to the optimal speculation point for each different batch size.

We make four key conclusions. First, combining batching with speculative decoding provides performance improvements. For example, in Figure 1b with batch size 4, having speculation length 3 improves speed by 1.93 $\times$ . Second, we observe that the benefits of speculative decoding are high when using small batch sizes, i.e., increasing the speculation length improves the per-token latency in smaller batch sizes. When the batch size is 1, speculative decoding reduces the per-token latency by up to 63%. This is because GPU utilization is low with small batch size thus using higher speculation length helps. Third, we find that when using large batch sizes (e.g., 16 or 32), higher speculation lengths incur performance slowdowns. When the batch size is 32, the smallest per-token latency is achieved using a speculation length smaller or equal to 2. This is because large batch sizes already fully utilize the underlying GPU computational resources, thus improving the execution parallelism with speculative decoding cannot provide further benefits. Fourth, the best-performing speculation length depends on the batch size and the characteristics of GPU architecture and LLM model. With a batch size of 8 in RTX3090, the optimal speculation length is 3 for OPT-1.3B and 4 for OPT-6.7B model. Similarly, using batch size 4 in OPT-6.7B execution, the optimal speculation length is 2 for A100 and 3 for RTX 4090. We conclude that an adaptive policy is necessary to autotune the speculation length based on the bathe size and the characteristics of LLM model and GPU architecture.

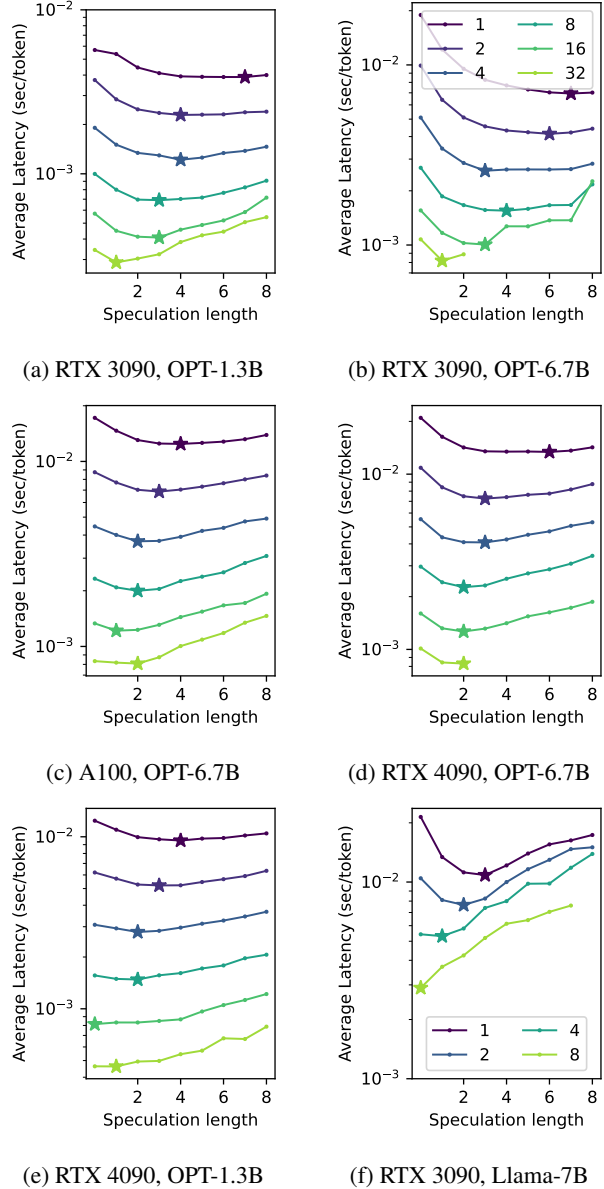


Figure 1. How per token latency changes with the batch size and speculation length, for different models and GPUs. The optimal speculation length is highlighted. A common phenomenon is that the optimal speculation length is smaller for larger batch sizes.

### 3.3. Modeling Analysis

To analyze the reason behind the aforementioned LLM behavior, we model the total execution runtime of batched LLM inference with speculative decoding. Table 1 lists the notations that we use in our modeling.

Intuitively, the total execution runtime with speculative decoding consists of two parts: (i) the runtime of LLM ( $T_L$ ), and (ii) the runtime of the SSM ( $T_S$ ).

Table 1. Notations used for modeling the LLM execution runtime.

$N$	The total number of tokens to be generated
$b$	The batch size to be used in inference
$s$	The number of speculated tokens generated at each speculative step
$l(s)$	Average number of correct tokens generated by the SSM
$T_L(b, s)$	Total execution time of the LLM
$T_S(b, s)$	Total execution time of the SSM
$t_L(b, s)$	Per step runtime of the LLM with batchsize= $b$ and query length $s$
$t_S(b, s)$	Runtime of the SSM with batchsize= $b$ and query length $s$

Both  $T_L$  and  $T_S$  depend on the batch size  $b$  and the speculation length  $s$ . Thus, we formulate the total execution runtime as follows:

$$T_{total} = T_L(b, s) + T_S(b, s) \quad (1)$$

For simplicity, we only consider the runtime of the auto-regression stage, neglecting the prefilling stage (i.e. the first iteration where the SSM and LLM reads the whole prompt). We also adopt the assumption that each iteration of these two models takes the same time. This is true when the context is not extremely long, and the bulk of the runtime is spent on matrix multiplications other than attention, which is proportional to the input query length. During the auto-regression stage, the length of the input query is fixed.

Based on these assumptions, we can write the total runtime of the LLM and the SSM as the product of the number of calls to them and the runtime of each call as follows. Used notations are listed in Table 1. As mentioned in Algorithm 1, in each iteration, the LLM always provides one extra correct token than  $l(s)$ . Therefore, the number of iterations is  $\frac{N}{l(s)+1}$ . We can thus write the runtime of the LLM and SSM as:

$$T_L(b, s) = \frac{N}{l(s)+1} \cdot t_L(b, s) \quad (2)$$

$$T_S(b, s) = \frac{N}{l(s)+1} \cdot (s \cdot t_S(b, 1)) \quad (3)$$

### 3.3.1. LLM Runtime $T_L$

LLM is used to verify the tokens generated by the SSM, and generates a new token to either correct wrong speculations or look one token ahead if the speculation is correct. Assuming  $l(s)$  provides the average number of correct tokens generated by SSM, the total number of tokens generated in the iteration is  $l(s) + 1$ . Thus, given that  $N$  is the number of token to be generated and  $t_L(B, s)$  is the execution time

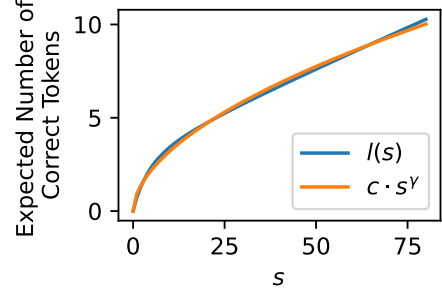


Figure 2. How  $l(s)$  scales with  $s$ , which can be approximated by a sublinear power function  $c \cdot s^\gamma$ , where  $\gamma < 1$ . The orange curve is an approximation  $0.9s^{0.548}$ .

for LLM to generate a token, the total LLM runtime can be estimated as  $\frac{N}{l(s)+1} \cdot t_L(B, s)$ .

### Expected Number of Correct Tokens $l(s)$

The expected number of correct tokens  $l(s)$  represents how accurately the SSM can mimic the behavior of the LLM. While it is hard to directly model it with a closed form, we did an empirical measurement to observe how it scales with  $s$ . We sampled  $n = 200$  prompts as inputs and let the SSM generate  $m = 80$  tokens for each prompt. The generated tokens are verified by LLM and the numbers of correct tokens are recorded. We denote the number of correct tokens generated for the  $i$ -th prompt as  $l_i$ . For an  $s \leq 80$ , we can use  $l_i$ -s to approximate  $l(s)$ , the average number of correct tokens in the first  $s$  tokens, as follows:

$$l(s) \approx \frac{1}{n} \sum_{i=1}^n \min\{l_i, s\} \quad (4)$$

The measured  $l(s)$  curve is shown in the blue curve in Figure 2. We can use a power function

$$l(s) \approx c \cdot s^\gamma \quad (5)$$

to approximate it. In this case, the approximation function is  $0.9s^{0.548}$ , which is plotted as the orange curve.

This is also consistent with the intuition that  $l(s)$  is a non-decreasing and sub-linear function ( $\gamma$  is always smaller than 1). This is because we can decompose  $l(s)$  as following:

$$l(s) = \sum_{i=1}^s p\left(\bigwedge_{j=1}^i E_j\right) \quad (6)$$

where  $p$  means probability and  $E_j$  is the random event that the  $j$ -th token is correct. Every term in the summation is non-negative. As  $s$  increases, it is harder to make correct speculation, so the probability  $p$  will decrease rapidly.

### LLM Runtime for Each Speculation Step $t_L(b, s)$

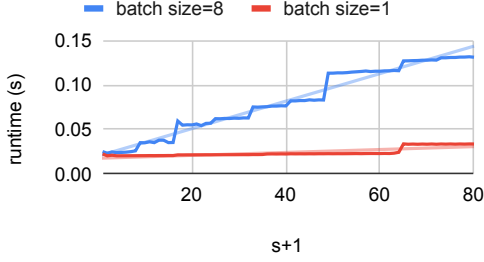


Figure 3. Execution time  $t_L(B, s)$  achieved when varying the speculation length  $s$  for different batch sizes  $B$ -s, which can be approximated by a linear function.

The runtime of one verification step of LLM  $t_L(b, s)$  depends on the hardware and model implementations. To illustrate how it scales with  $s$ , we measured the runtime of OPT-6.7B on an RTX 3090 GPU, and the result is shown in Figure 3. The runtime curves are approximately step functions. The first jump of the curve of  $b = 1$  is when  $s = 64$ , while that of  $B = 8$  jumps up much earlier when  $s = 8$ . For simplicity, we approximate  $t_L(b, s)$  as linear functions with different slopes  $t_L(b, s) = \alpha_b s + \beta$ .

### 3.3.2. SSM Runtime $T_S$

In practice, we usually also need to consider the second term, i.e. the runtime of the small model. For a fixed  $B$ ,  $t_S(B, 2)$  is a constant, and the constant is also increasing with respect to  $B$ .

### 3.3.3. Optimal Speculation Length $s_{opt}$

Based on the aforementioned analysis, the total runtime can be written as

$$t = \frac{N \cdot t_L(b, s)}{l(s) + 1} + \frac{N \cdot s \cdot t_S(b, 1)}{l(s) + 1} \quad (7)$$

$$\approx N \frac{\alpha_b \cdot s + \beta + t_S(b, 1) \cdot s}{c \cdot s^\gamma + 1} \quad (8)$$

Since  $t_S(b, 1)$  is a constant regarding  $s$ , and is also increasing with  $b$  as  $\alpha_b$ , we merge it with  $\alpha_b$  for simplicity.

The minimum point of this function  $s_{opt}$  is where the derivative of  $t$  is zero, i.e.  $\frac{\partial t}{\partial s} = 0$ . And we have

$$\frac{\partial t}{\partial s} = N \frac{\alpha_b (c \cdot s^\gamma + 1) - (\alpha_b \cdot s + \beta) \cdot c \gamma s^{\gamma-1}}{(c \cdot s^\gamma + 1)^2} \quad (9)$$

It is zero when the numerator is zero, which can be written as

$$(1 - \gamma) \alpha_b \cdot c s^\gamma - c \cdot \beta \cdot \gamma s^{\gamma-1} + \alpha_b \quad (10)$$

Let  $K = (1 - \gamma) \cdot c$ ,  $L = c \cdot \beta \cdot \gamma$ , we can rewrite the expression above in the following form:

$$\delta = K \alpha_b \cdot s^\gamma - L s^{\gamma-1} + \alpha_b \quad (11)$$

The optimal speculation length  $s_{opt}$  needs to satisfy  $\delta = 0$ .

$$K \alpha_b \cdot s^\gamma - L s^{\gamma-1} + \alpha_b = 0 \quad (12)$$

Since  $K \cdot s^\gamma > 0$ ,  $L s^{\gamma-1}$  is a constant regarding  $b$ , and  $\alpha_b$  is increasing with  $b$ ,  $\delta$  is monotonically increasing with respect to  $b$ . Similarly, because  $K \cdot \alpha_b > 0$  and  $\gamma - 1 < 0$ ,  $\delta$  is also increasing with  $s$ . Therefore, for a larger  $b$ ,  $s_{opt}$  needs to be smaller in order to keep the  $\delta$  as zero, and vice versa, i.e., for a smaller  $b$ ,  $s_{opt}$  needs to be larger.

## 4. Adaptive Speculative Decoding: Our Approach

In real-world serving scenarios, the traffic might be time-varying, leading to varying batch sizes. Based on the observation and analysis that the optimal speculation length depends on the batch size, we propose a new adaptive speculation mechanism to choose the optimal speculation length under different scenarios.

The key point is to find the optimal speculation step  $s_{opt}$  for each different batch size  $b$ . While it is hard to find a closed-form solution for it, we can empirically measure the and do a grid search to find  $s_{opt}$ . We divide the process into two stages, namely profiling and execution. In the profiling stage, we use a small sample of the dataset (similar to the training set in machine learning) to measure the latency under different batch sizes using different speculation lengths. Using the profiling result, we can build a look-up table (LUT) that stores the optimal speculation length for each batch size. During the execution, after a batch of requests arrives, we can find the optimal speculation length using the LUT, and configure the LLM execution, accordingly.

The overhead of profiling is negligible in LLM serving scenarios for two key reasons. First, we enable a very small search space. The optimal speculation length is usually small (less than ten). While profiling all possible batch sizes can be prohibitive. Instead, we sample only a subset of them. We profile batch sizes which are powers of two (1, 2, 4, 8, 16, ...). For batch sizes that are not profiled, we choose the smaller speculation length of the nearest two profiled batch sizes. Second, the profiling overhead can be amortized, because LLM serving is usually a long-running serving task. A service can be online for weeks or months serving millions or even billions of user requests, while profiling only takes several minutes and it is only done once: before launching the LLM inference process in the server.

## 5. Evaluation

We evaluate the performance benefits of our proposed adaptive batched speculative mechanism using the methodology

presented in Section 5.1 in two aspects: (i) quantifying a uniform traffic, where user requests are served by LLM at a fixed batch size (Section 5.2), and (ii) quantifying a dynamic traffic, where user requests are served by LLM at a time variable batch size (Section 5.3).

### 5.1. Methodology

**Dataset.** We use the *Chatbot Instruction Prompts* dataset (Palla, 2023) from Huggingface. We only use the prompt field of it as the input.

**Hardware and Software.** We use NVIDIA RTX 3090 as our benchmark platform. We use Ubuntu 22.04, CUDA 11.8, and PyTorch 2.0.1.

**Models.** We use OPT-6.7B (Zhang et al., 2022) as our large language model (LLM) and OPT-125M as the small speculative model (SSM).

### 5.2. Uniform Traffic

In this section, we assume that user requests are served by LLM with a fixed batch size. We compare our approach, i.e., adaptive speculation decoding, that integrates using the optimal speculation length, over the baseline LLM serving without speculation decoding. We sample 1000 prompts from the dataset and group them into batches. For each prompt, we generate 128 new tokens. Figure 4 presents the end-to-end time of finishing all prompts, including the tokenizer, by normalizing the results over the baseline without speculation decoding. We did a grid search to find the optimal speculation length.

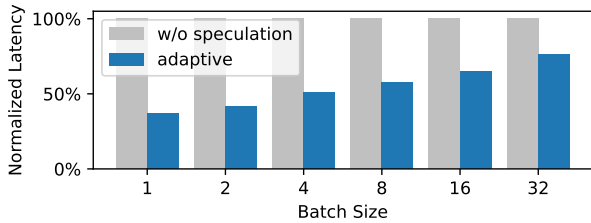


Figure 4. Normalized latency with fixed batch sizes.

We make two observations. First, adaptive speculative decoding provides higher performance benefits for smaller batch sizes. For example, when the batch size is 1, adaptive speculating decoding achieves a  $2.73\times$  speedup (63% latency reduction), while when the batch size is 32, the speedup is  $1.31\times$ . This is because when batch size is small GPUs are highly underutilized, and it is leveraged by speculative decoding. When batch size is large, GPUs are well utilized, so the room for performance improvement is limited. Second, across all batch sizes, adaptive speculative

decoding provides a high speedup by on average of  $1.94\times$  over LLM serving with no speculative decoding.

### 5.3. Dynamic Traffic

In this section, we measure the effectiveness of adaptive speculative decoding in a real-world scenario, where the traffic highly varies over time. We create a server-client setting, and we artificially vary the number of requests sent to the LLM server per second.

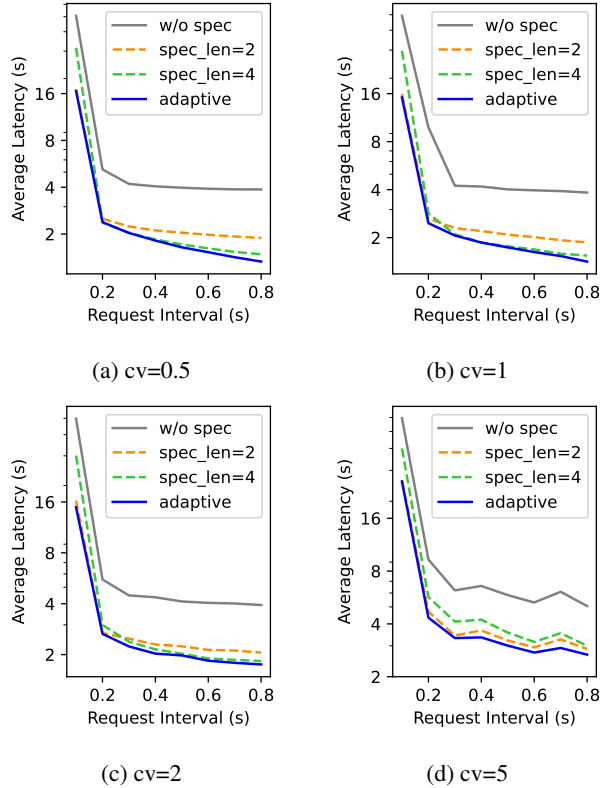


Figure 5. Average latency with various traffic scenarios, i.e., varying request intervals and coefficient of variation (CV).

**Client Configuration.** We launch a client process that samples a sequence of requests from the dataset and sends them to the server at a fixed interval time. The interval time is sampled from a Gamma distribution, and the parameters of the Gamma distribution allow us to control the average interval between requests as well as the coefficient of variation (CV). This way we can control the overall volume of traffic by varying the average interval, and the variation and blast of the traffic by varying the CV.

**Server Configuration.** We launch a server process and wrap the LLM inference as a service that receives requests from a message queue and responds to the generated tokens via another message queue. If there is more than one request

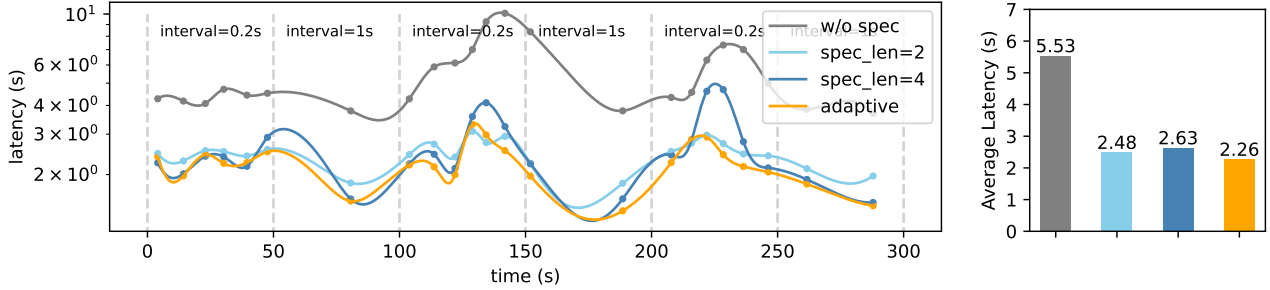


Figure 6. Timeline of four speculation strategies. Each point represents a group of 40 requests. The X-axis is the time stamp of the first request in a group. The Y-axis is the average latency of all requests in a group of requests. There is no optimal speculation length for all scenarios. When the request traffic is intense (e.g. 100-150, 200-250), a smaller speculation length like 2 is preferred. When the request traffic is sparse (e.g. 150-200, 250-300), a larger speculation length like 4 is preferred. Nevertheless, adaptive speculation decoding can always achieve performance on par with or better than both fixed speculation length schemes.

in the queue, they will be merged as one batched request (up to a maximal batch size of 16 due to the memory constraint).

**Comparison Points.** We compare our proposed adaptive speculative decoding scheme with three schemes: 1) Batched LLM inference without speculative decoding, 2) Batched LLM inference with speculative decoding of fixed speculation length equal to 2, and 3) Batched LLM inference with speculative decoding of fixed speculation length. For fixed speculative decoding schemes, we select speculation lengths of 2 and 4, since they achieve better performance. We do not evaluate the speculative decoding scheme without batching, because it performs significantly worse than our remaining evaluated schemes.

**Experiments.** We create dynamic request traffic on the client side and evaluate various volume and blast traffic configurations using four different values for CV (0.5, 1, 2, and 5) and eight different values for the interval time from 0.1s up to 0.8s with a step of 0.1s. For each setting, we generate only one sequence of requests, which is used to evaluate all comparison points. Each ‘sequence’ contains 1000 prompts, and there are no overlaps between the dataset used in the profiling step of our proposed scheme and the dataset used in our dynamic traffic evaluation. We record the time  $t_a$  at which each request is sent by the client to the server and the time  $t_b$  at which the server finishes serving the request and measure the latency as the difference  $t_b - t_a$ . Note that if the server is slow, requests cannot be processed in time and have to wait in a queue, thus the waiting time is also accounted for in the latency. Figure 5 presents the average latency across all requests achieved by all comparison points using different configurations for dynamic traffic generation.

We draw two findings. First, adaptive speculative decoding can provide high latency benefits in all various scenarios, including intense (low request interval) and sparse (high

request interval) traffic, as well as uniform (low CV) and skewed (high CV) traffic scenarios. On average, adaptive speculative decoding achieves a speedup of  $2.3\times$  over LLM serving with no speculative decoding. Second, we observe that the best-performing speculation length varies across different traffic scenarios, i.e., there is no speculation length value that provides the lowest latency across all traffic scenarios. For example, when having  $CV=2$  and a request interval equal to 0.1, a speculation length of 2 achieves the best latency, while when having  $CV=2$  and request interval equal to 0.6, a speculation length of 4 achieves the best latency. We observe that our proposed scheme effectively adapts the speculation length to the characteristics of the current traffic and achieves the best performance, outperforming both fixed speculation decoding schemes. Especially when the variation of the traffic is large (e.g. when CV is equal to 5), because the batch size might vary with the traffic volume, adaptive speculative decoding provides an additional  $1.07\times$  average speedup over the better one of the two fixed speculation length. The speedup over fixed speculation length can be up to  $1.15\times$ .

Figure 6 presents the timeline on latency achieved by all schemes during runtime under time-varying traffic scenarios. In this experiment, the client alternates between two types of traffic: *intense* with a request interval of 0.2 seconds and *sparse* with a request interval of 1 second. The client will switch from one type of traffic to the other every 50 seconds. We fixed the CV as 1 in this experiment.

We find that adaptive speculation decoding effectively adapts to best-performing configuration. A speculation length of 4 performs better for sparse request traffic because the batch size is usually small, and a speculation length of 2 is more suitable for intense traffic. For example, from the 150th to 200th second, the traffic is sparse, therefore the speculation length of 4 performs better than the speculation

length of 2. The situation is opposite from the 200th to 250th seconds. Nevertheless, adaptive speculation decoding can always match the performance of the better one of the two static baselines, or even perform better. Overall, adaptive improves latency over speculation lengths of 2 and 4 by 9% and 14%, respectively, on average. We conclude that our approach provides significant performance benefits by dynamically tuning itself based on the characteristics of the current traffic and leveraging the best of both speculative decoding and batching optimizations.

## 6. Conclusion

LLMs like GPT achieve significant success nowadays. However, LLM execution is inherently sequential, thus incurring low hardware utilization on GPUs. Batching and speculative decoding are two techniques to improve the GPU hardware utilization in LLM inference. To study the synergy between them, we carry out a comprehensive study, and observe that the optimal speculation length depends on the batch size as well as the characteristics of LLM model and GPU architecture. We build a quantitative model to explain this key observation, and based on our analysis, we propose an efficient adaptive speculative decoding strategy that selects the optimal speculation length for different batch sizes. Our evaluations show that our proposed method achieves equal or better performance than fixed speculation decoding schemes, and we hope our work will enable further research on accelerating LLM serving.

## References

- Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choremanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- 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., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Crankshaw, D., Wang, X., Zhou, G., Franklin, M. J., Gonzalez, J. E., and Stoica, I. Clipper: A {Low-Latency} online prediction serving system. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pp. 613–627, 2017.
- Fang, J., Yu, Y., Zhao, C., and Zhou, J. Turbotransformers: an efficient gpu serving system for transformer models. In *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pp. 389–402, 2021.
- Gao, P., Yu, L., Wu, Y., and Li, J. Low latency rnn inference with cellular batching. In *Proceedings of the Thirteenth EuroSys Conference*, pp. 1–15, 2018.
- Ivanov, A., Dryden, N., Ben-Nun, T., Li, S., and Hoefler, T. Data movement is all you need: A case study on optimizing transformers. *Proceedings of Machine Learning and Systems*, 3:711–732, 2021.
- Lee, Y., Scolari, A., Chun, B.-G., Santambrogio, M. D., Weimer, M., and Interlandi, M. {PRETZEL}: Opening the black box of machine learning prediction serving systems. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pp. 611–626, 2018.
- 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.
- Liu, X., Hu, L., Bailis, P., Stoica, I., Deng, Z., Cheung, A., and Zhang, H. Online speculative decoding. *arXiv preprint arXiv:2310.07177*, 2023.
- Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Wong, R. Y. Y., Chen, Z., Arfeen, D., Abhyankar, R., and Jia, Z. Specinfer: Accelerating generative llm serving with speculative inference and token tree verification. *arXiv preprint arXiv:2305.09781*, 2023.
- Olston, C., Fiedel, N., Gorovoy, K., Harmsen, J., Lao, L., Li, F., Rajashekhar, V., Ramesh, S., and Soyke, J. Tensorflow-serving: Flexible, high-performance ml serving. *arXiv preprint arXiv:1712.06139*, 2017.
- OpenAI. Gpt-4 technical report, 2023.
- Palla, A. Chatbot instruction prompts datasets. [https://huggingface.co/datasets/alespalla/chatbot\\_instruction\\_prompts](https://huggingface.co/datasets/alespalla/chatbot_instruction_prompts), 2023.
- Shen, H., Chen, L., Jin, Y., Zhao, L., Kong, B., Philipose, M., Krishnamurthy, A., and Sundaram, R. Nexus: A gpu cluster engine for accelerating dnn-based video analysis. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pp. 322–337, 2019.

- Spector, B. and Re, C. Accelerating llm inference with staged speculative decoding. *arXiv preprint arXiv:2308.04623*, 2023.
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- 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 Lample, G. Llama: Open and efficient foundation language models, 2023.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- 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 *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pp. 521–538, 2022.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.