DeltaZip: Efficient Serving of Multiple Full-Model-Tuned LLMs

Xiaozhe Yao
ETH Zurich

Qinghao Hu
MIT

Ana Klimovic
ETH Zurich

Abstract

Fine-tuning large language models (LLMs) greatly improves model quality for downstream tasks. However, serving many fine-tuned LLMs concurrently is challenging due to the sporadic, bursty, and varying request patterns of different LLMs. To bridge this gap, we present DeltaZip, an LLM serving system that efficiently serves multiple full-parameter fine-tuned models concurrently by aggressively compressing model deltas by up to 10× while maintaining high model quality. The key insight behind this design is that fine-tuning results in small-magnitude changes to the pre-trained model. By co-designing the serving system with the compression algorithm, DeltaZip achieves 2× to 12× improvement in throughput compared to the state-of-the-art systems.

1 Introduction

Large Language Models (LLMs), such as GPT [51], Llama [64] and Gemini [63] have demonstrated remarkable performance and are widely used in a variety of applications, such as chatbots [49] and coding assistants [14, 28, 84]. To achieve high accuracy for a target domain, LLMs are first pre-trained on a large corpus of text data, then fine-tuned on application-specific tasks or datasets [43, 56], such as code [14], conversations [52], and human preferences [55]. Cloud AI infrastructure companies, such as OpenAI [48], Google [54], Microsoft [44], and Anyscale [5] expose APIs for users to fine-tune a pre-trained LLM with their own data and deploy the resulting customized model variant for inference.

While fine-tuning is typically a one-time effort performed off the critical path, LLM serving is critical to optimize as it is typically recurring and latency-critical. Techniques such as continuous batching [76], paged attention [36], prompt processing disaggregation [53, 61, 82], and tensor parallelism [39] optimize inference latency and throughput for individual models, however concurrently serving many model variants presents additional challenges.

Figure 1 shows that the requests to each model variant in an LLM-based chatbot service are sporadic, making it difficult to batch requests and decide how to provision GPU resources for each model. Dedicating GPUs for each model variant minimizes latency, but requires many expensive GPUs that would sit idle for long periods of time (yellow areas in Figure 1). On the other hand, swapping model variants on a limited pool of GPUs reduces cost and improves utilization, but adds latency on the critical path of requests. Optimizing the storage hierarchy to reduce swapping latency can

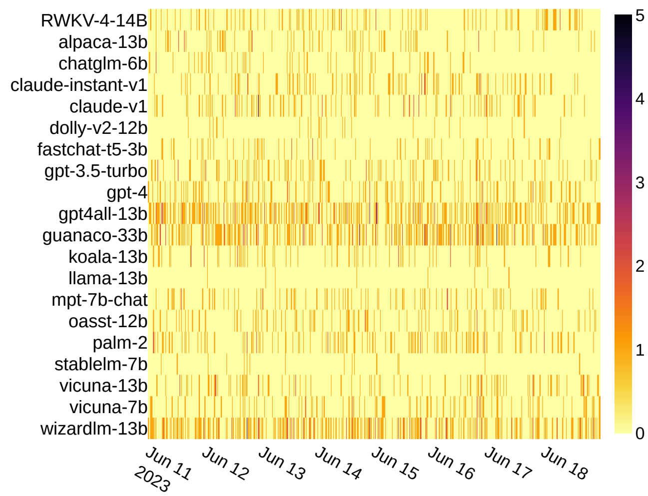

Figure 1. Invocation counts per 5-min time windows for 20 different models in the LMSys Chatbot Arena [81] trace.

help [26], but requests still experience long queuing delays due to limited batching opportunities when treating each model variant as a separate model.

The state-of-the-art approach to LLM serving with the proliferation of fine-tuned model variants is to adopt an entirely new fine-tuning paradigm, *parameter-efficient tuning* (PEFT). Instead of tuning model weights directly, PEFT tunes a compact, adjunctive model adapter (i.e., a small set of extra parameters) that is attached to the model for serving. For example, low-rank adaptation (LoRA) [32] is a popular PEFT method that freezes models weights and attaches low-rank matrices to the model structure which are fine-tuned on task-specific data. Systems like Punica [13] and S-LoRA [59] leverage the small size of LoRA adapters and the common base model weights to efficiently swap adapters and batch requests to optimize inference latency and throughput.

However, PEFT serving systems are not compatible with the traditional fine-tuning approach, *full model tuning* (FMT). While PEFT methods have achieved high accuracy for downstream tasks like SQL generation [5, 7] and ViGGO [35], they are still not able to match the accuracy of FMT for more complex tasks, such as coding and math [9], or when the fine-tuning dataset is particularly large [78]. Figure 2 summarizes the results of two recent studies [5, 9] comparing LoRA and FMT accuracy on three downstream tasks. Full-parameter fine-tuning remains appealing to applications that aim to maximize accuracy on more complex tasks and is still widely-used. Yet the serving solutions available for this paradigm

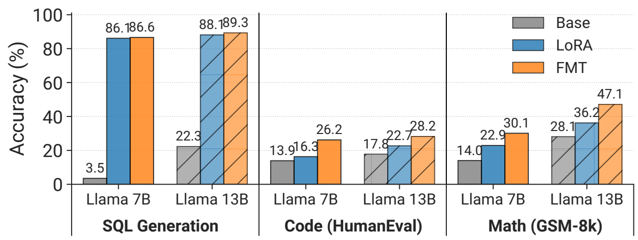

Figure 2. LoRA vs. full-model fine-tuning accuracy [5, 9]. LoRA fine-tuning is comparable for some tasks (SQL), but has lower quality on more complex tasks (Math and Code).

(which either dedicate GPUs for each model variant or swap entire models) are either expensive or slow.

We aim to extend PEFT-based serving systems to also support efficient FMT model variant serving. Our key insight is that FMT model weights often have low-magnitude perturbations with respect to the original pre-trained model (see Figure 3), allowing us to aggressively sparsify, quantize, and compress *model deltas* while maintaining high accuracy. Due to their compact size, low-precision and sparse deltas can be swapped and served with low latency. We apply this idea in the design of DeltaZip, a multi-variant LLM serving system. DeltaZip extracts and compresses model deltas with ΔCompress, an algorithm we propose to help maintain high accuracy during compression. To efficiently serve FMT model variants, DeltaZip decouples base model serving and low-precision delta serving, inspired by how S-LoRA [59] and Punica [13] serve LoRA adapters. This decoupling enables DeltaZip to batch requests to different model variants that share the same base and perform low-precision inference for model deltas to minimize latency and memory bandwidth pressure on the GPU. We further optimize delta inference by designing a custom GPU kernel, Selective Batched Matrix Multiplication (*SBMM*), which selectively batches requests to the same delta to minimize random accesses and performs operations for multiple deltas in parallel to amortize kernel launch overhead. We build DeltaZip on top of vLLM [36] (which supports S-LoRA/Punica-based LoRA-serving [12, 59]) and adapt continuous batching [76] and model parallelism [60] for delta-based model serving.

To the best of our knowledge, DeltaZip is the first serving system to support both FMT and PEFT model variants while accelerating FMT model serving with hardware-optimized delta compression. In summary, our key contributions are:

- We propose ΔCompress (§4), a *hardware-efficient* compression algorithm that aggressively compresses model deltas post full-model fine-tuning. It applies structured sparsity, quantization, and optionally lossless compression. ΔCompress can compress a 70B-parameter Llama-2 model delta by 13× while maintaining comparable accuracy to the original FMT model. In contrast, applying a similar sparsification and quantization technique like SparseGPT [23] directly on the fine-tuned model substantially degrades accuracy, even with only 6× compression.
- We design DeltaZip (§5), a serving system that leverages ΔCompress to efficiently serve FMT model variants. It decouples and parallelizes base and delta model inference, reducing queuing delays by batching requests to different model variants derived from the same base and leveraging our custom GPU kernel for hardware-optimized low-precision and sparse delta serving. DeltaZip achieves 2× to 12× higher throughput compared to vLLM [36].
- We identify challenges when deploying DeltaZip in practice and propose solutions to address them. In particular, we study how many deltas to place concurrently to balance GPU memory usage and how to reduce starvation when serving many model deltas.

2 Background and Motivation

We first introduce key concepts in LLM fine-tuning and serving, then highlight key challenges and opportunities.

2.1 Background

LLM fine-tuning. In contrast to prior deep learning workloads that follow a *task-specific* paradigm, which trains a model from scratch on domain-specific data to tackle a particular task (e.g., machine translation [62, 71]), LLMs follow a *pretrain-then-finetune* paradigm, which pretrains a model on a massive general-domain corpus and then fine-tunes it for specific objectives. For example, ChatGPT is fine-tuned to follow human instructions [52]. There are two main fine-tuning approaches: 1) *Full Model Tuning* (FMT) which updates all model parameters and 2) *Parameter-Efficient Tuning* (PEFT) which adds a small number of extra parameters after pretraining, called *adapters*, e.g., low-rank matrices learned during fine-tuning. PEFT methods, such as LoRA [32], are popular ways to reduce the compute and memory requirements of both fine-tuning and serving. However, the choice of fine-tuning paradigm impacts accuracy. While PEFT methods can achieve high accuracy for a variety of tasks [32], recent studies [5, 9, 78] — summarized in Figure 2 — reveal that FMT still achieves higher accuracy for more complex tasks.

LLM compression. Model compression is a popular approach to reduce the memory and compute requirements of LLM inference in resource-constrained environments. Techniques like GPTQ [25], SparseGPT [23] and AWQ [40] reduce memory footprints and improve latency while maintaining model quality (when applied in moderation). Pushing these techniques to extremely low bit-width quantization and sparsity, such as 2-bit quantization and more than 50% sparsity, results in significant model quality degradation [11, 42].

LLM serving. The LLM inference involves two phases: (1) *prompt processing* (prefill), where the tokens (i.e., basic units of text) in the input prompt are processed in parallel. This phase can be parallelized since all previous tokens are known from the user-provided prompt, and it is usually compute-bound. (2) *token generation* (decode), where the model iteratively generates one token for each forward pass. Due to the inter-token data dependency, this phase cannot be parallelized and is typically memory-bound. Token generation stops when the model generates an end-of-sequence (EOS) token or meets a user-defined condition (e.g., reach the maximum number of generated tokens). Many works [13, 36, 76], including our own, focus on optimizing the token generation phase as it is the main bottleneck for LLM serving.

2.2 Challenges for Multi-Variant Model Serving

With the proliferation of fine-tuned LLM variants, each specialized for a particular user's task or domain, it is critical to design LLM serving systems that can efficiently multiplex requests to different model variants. Yet, on today's commercial LLM serving platforms, fine-tuned model variant serving is still more expensive than base model serving [1, 50]. The high cost is due to several key challenges that lead to low resource utilization:

Low request rates per model variant. A naive approach to serve multiple fine-tuned models is to consider each model as a separate model and dedicate a group of GPUs for each variant. However, as shown in Figure 1, the requests to each model variant are sporadic and often low in volume, which limits the batch size. Hence, dedicating GPUs for each model underutilizes resources and leads to high cost.

Swapping incurs high latency. Another approach is to swap the models in and out of a limited pool of GPUs. However, the unpredictable nature of model invocations makes it hard to predict when a request for a particular model will arrive to enable prefetching. Swapping models in and out of GPUs on the critical path of requests leads to high latency and GPUs also remain underutilized.

Accuracy gap between PEFT and FMT. Although PEFT methods such as LoRA [32] produce significantly smaller fine-tuned adapters and systems [12, 59, 70, 85] have been built for serving adapters, these methods face the challenge of model quality. As shown by recent studies [5, 9, 78] (summarized in Figure 2), FMT still achieves higher accuracy for more complex tasks compared to PEFT methods.

3 DeltaZip Overview

We observe three key opportunities to address the aforementioned challenges and improve the efficiency of FMT serving (§3.1). We then describe how we leverage these opportunities to design DeltaZip, a multi-variant LLM serving system that complements parameter-efficient model serving with hardware-efficient full model tuning serving (§3.2).

3.1 Opportunities and Key Insights

1. Model deltas are highly compressible. Although fine-tuning can significantly improve model performance on specific tasks, we find it typically results in small-magnitude

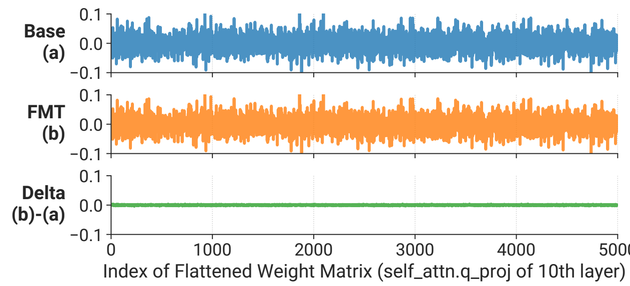

Figure 3. Flattened weight matrix in an intermediate layer of the pre-trained model (a), the fine-tuned model (b), and the model delta between the two (bottom, (b)-(a)).

changes to model parameters. Figure 3 shows the distribution of the weight matrix $\mathbf{w_q}$ in a transformer layer of a pre-trained `Llama-2-7b` model, its fine-tuned counterpart `Vicuna-7b-v1.5` model, and the model delta, which is obtained by subtracting the base model from the fine-tuned model. The delta has significantly smaller magnitude range and fewer outliers. This makes the delta easier to compress than the model itself, with both quantization [25, 40] and sparsification [23]. As quantization involves calculating the maximum value such that the quantization preserves the outliers in the weight matrix, a more concentrated distribution results in a denser quantization grid, which can maintain model quality while using lower bit width. In addition, the prevalence of near-zero values makes it easier to apply sparsification to the model delta than to the full model. We will show that this enables us to compress model deltas by over 10×, making it fast to swap and serve while achieving comparable quality to full-precision model serving.

2. Many model variants share the same base model, for which we can batch many requests. Since pre-training LLMs requires immense datasets and compute resources, only a handful of organizations are well-positioned to produce high-quality pre-trained models [3, 63]. Instead, most organizations rely on fine-tuning pre-trained models for their use cases, as this is typically much cheaper [1]. This means that fine-tuned models often share the same base model, even if they are used for different applications. For example, GitHub's Copilot [19] and OpenAI's ChatGPT [52] are both fine-tuned from GPT models. Hence, although batch sizes for individual model variants may be limited (Figure 1), we can quickly accumulate large batches to common base models to improve GPU utilization and reduce queuing delays. This requires decoupling base model and delta serving, which we will discuss in §5.1.

3. GPU support for delta computations. We can accelerate delta inference by leveraging features of modern GPU hardware, such as: 1) using sparse tensor cores [8] for sparse delta computations, 3) performing multiple low-precision matrix multiplications in parallel to improve stream multiprocessors (SMs) utilization, and 3) reducing memory bandwidth pressure for delta matrix computations by designing

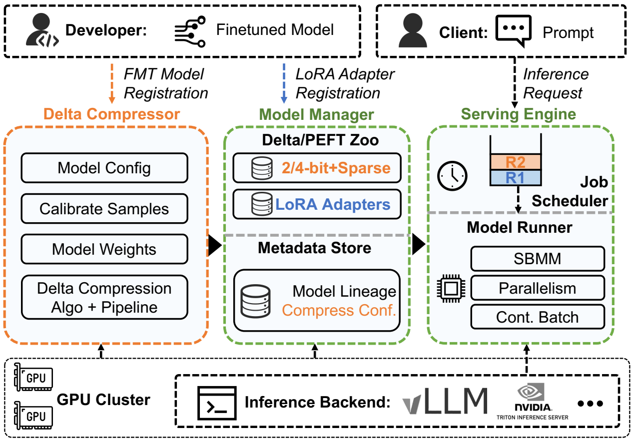

Figure 4. DeltaZip system architecture.

a custom GPU kernel that minimizes data movement from GPU global memory to device memory.

3.2 System Overview

To realize the above opportunities in practice, we propose DeltaZip. Our key contributions are a model delta compression algorithm (Δ*Compress*) and the end-to-end DeltaZip serving system. The algorithm compresses FMT deltas into a hardware-efficient, low-precision, and sparse format that preserves high accuracy, while the serving system incorporates state-of-the-art serving optimizations and adapts them for low-latency, high-throughput inference with model deltas.

System architecture. Figure 4 shows the DeltaZip system architecture. The system consists of three main components: 1) the *Delta Compressor* (§4) which extracts and compresses the delta from a FMT model registered by the user, 2) the *Model Manager* which tracks and manages model metadata, and 3) the *Serving Engine* (§5) which serves inference requests for base and fine-tuned models.

Life of a request. The model developer first uploads a fine-tuned model to the Delta Compressor, together with some metadata (such as the pre-trained model identifier) and a small calibration dataset that the compression algorithm uses to measure and minimize accuracy loss. The compressor computes and compresses the model delta, then stores it in a packed, low-precision, and sparse format in the Model Manager's delta zoo. The manager keeps track of metadata for each stored delta, such as its compression configuration (such as the bit width per parameter and sparsity level) and model lineage. In addition to deltas, the model manager also allows developers to register LoRA adapters directly, and the serving engine can serve them as well.

The Serving Engine serves inference requests for fine-tuned models whose deltas are stored in the model manager. Users send inference requests to the engine's API frontend, which fetches the requested deltas into CPU main memory, if they are not already present in CPU or GPU memory. Meanwhile, the frontend forwards the request to the job scheduler, which queues the requests. The job scheduler performs continuous batching by assigning requests to the model runner per-iteration (i.e., for each forward pass of the model). The model runner is responsible for processing the batched requests. Internally, the model runner can process requests for different FMT and PEFT models in parallel, by decoupling base model and delta (or PEFT adapter) inference computations. The model runner also leverages tensor parallelism and supports large models that do not fit in a single GPU.

4 Model Delta Compression

When a model developer registers an FMT model with DeltaZip, the Delta Compressor computes the delta and applies a compression pipeline (§4.1) to reduce the size of the model state. The compression pipeline only runs when a developer registers a new model. We develop the ΔCompress algorithm (§4.2) to apply quantization and sparsification in the compression pipeline in a way that preserves high accuracy.

We design DeltaZip to be agnostic to the compression pipeline, such that users can apply a variety of compression techniques to model deltas, such as GPTQ [25], SparseGPT [23], and AWQ [40]. Here we describe a SparseGPT-inspired compression pipeline, which we use in our implementation and evaluation of DeltaZip. We implemented this compression pipeline on top of AutoGPTQ [68] and SparseGPT [23].

4.1 DeltaZip Compression Pipeline

Within DeltaZip, the Delta Compressor encompasses multiple steps as shown in Figure 5. Initially, Step 1 computes the model delta by subtracting the pre-trained base model weights from the fine-tuned model. The remaining steps are applied on the delta matrix since it has a smaller range of values (see Figure 3) and is hence more amenable to compression compared to the fine-tuned model weight matrix.

Step 2 applies structured 2:4 pruning [33, 83], which involves setting at least two elements among each group of four contiguous elements to 0 in the delta matrix. After structured pruning, DeltaZip only needs to store the non-zero values of the delta matrix and their 2-bit indices. Compared to applying quantization only or unstructured sparsity, structured sparsity enables higher peak performance with large input size. Figure 6 shows a microbenchmark of matrix multiplication performance with different compression configurations. We observe that with small input sizes (e.g., from 1 to 4, which is common during the decode phase), structured sparse matrix multiplication achieves similar performance to quantization-only compression and outperforms the uncompressed version. This is mainly because both sparsity and quantization reduce data movement between GPU HBM and compute unit. However, with large input sizes (e.g., from 16 to 4096, which is typical during the prefill phase), structured sparsity significantly outperforms quantization-only compression. This is because structured sparsity leverages the sparse tensor cores on GPUs to achieve higher performance.

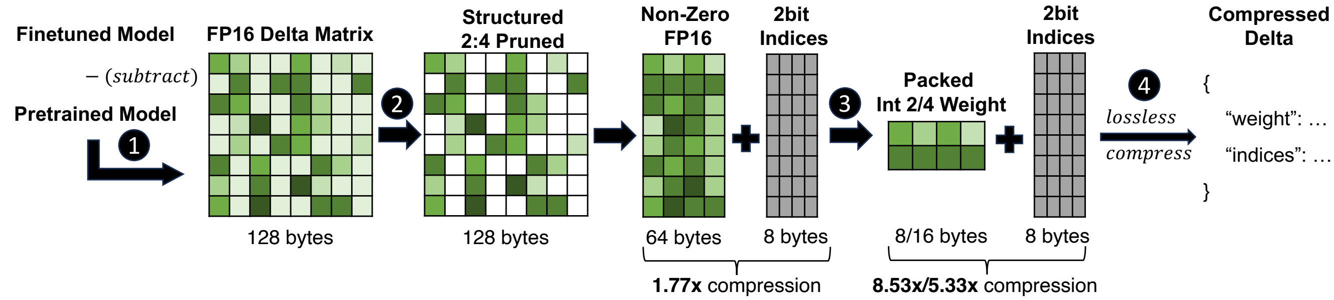

Figure 5. DeltaZip Compression Pipeline. The pipeline consists of delta extraction, sparsification & quantization, and optionally lossless compression. The compressed delta is stored as a dictionary of compressed weight matrices and metadata.

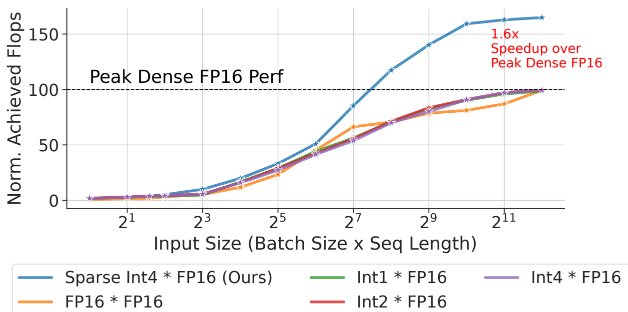

Figure 6. (Compressed) Matrix Multiplication Performance

Algorithm 1 ΔCompress algorithm. Given an N-layer FMT model where each layer has a weight matrix $\mathbf{w}_f$ of the shape $d_{\text{row}} \times d_{\text{col}}$, and the corresponding layer in the base model has a weight $\mathbf{w}_b$, the algorithm computes quantized weight $\mathbf{Q}$ and pruning mask $\mathbf{M}$. $\odot$ denotes elementwise multiplication.

1: **for** $n = 0, 1, 2, \cdots, N$ **do**
2: $\quad \mathbf{M} \leftarrow \mathbf{1}_{d_{\text{row}} \times d_{\text{col}}}$ ▷ *Binary Pruning Mask*
3: $\quad \mathbf{Q} \leftarrow \mathbf{0}_{d_{\text{row}} \times d_{\text{col}}}$ ▷ *Quantized Delta*
4: $\quad \Delta = \mathbf{w}_f - \mathbf{w}_b$ ▷ *Extract Delta*
5: $\quad \mathbf{Q}, \mathbf{M} = \text{Compress}(\Delta, \mathbf{X}_n)$ ▷ *e.g., SparseGPT*
6: $\quad \tilde{\mathbf{w}}_q \leftarrow \mathbf{Q} \odot \mathbf{M} + \mathbf{w}_b$ ▷ *Reconstruct Weight*
7: $\quad \mathbf{X}_{n+1} = \tilde{\mathbf{w}}_f \mathbf{X}_n$ as input of next layer.
8: **for** $n = 0, 1, 2 \cdots, N$ **do**
9: $\quad$ pack and store $\mathbf{Q}, \mathbf{M}$ of layer n.

Step 3 quantizes the pruned delta weight matrix, which squeezes the values into a smaller bit-width format and packs the data. For example, 4-bit quantization packs 8 values into a single 32-bit value, achieving 4× compression ratio.

Step 4 is an optional final step that applies lossless compression. We use the GDeflate algorithm from `nvcomp` [47] for lossless compression for fast decompression on GPUs. This step is beneficial when the disk bandwidth is a bottleneck (such as with NFS). In such cases, users can opt-in to lossless compression to reduce the disk I/O time. If disk I/O is not a bottleneck, lossless compression may not be beneficial due to the decompression overhead.

Hardware-efficient design. We design DeltaZip's compression pipeline with GPU hardware features in mind. Structured pruning (Step 2) leverages sparse tensor cores [8] for fast sparse matrix multiplication. The quantization (Step 3) allows us to move a smaller amount of data between GPU global memory to device memory, alleviating the memory bandwidth bottleneck. Lossless compression (Step 4) uses GPU decompression engines [47].

4.2 ΔCompress Algorithm

The core of the compression pipeline is the lossy compression algorithm that finds the optimal pruning mask (in Step 2) and quantized weight matrix (in Step 3). We design ΔCompress to compress the model deltas in a way that minimizes the loss between the outputs computed by the original weights and the compressed weights. ΔCompress achieves this by calibrating the compression algorithm with a subset of the training dataset, provided by model developers when registering the FMT model to DeltaZip. As outlined in Algorithm 1, ΔCompress iteratively (Line 1) compresses the delta (extracted in Line 4) for each layer of the model. For each layer, the objective is to find the optimal pruning mask $\mathbf{M}$ and quantized weight matrix $\mathbf{Q}$ (Line 2, Line 3). ΔCompress is designed to be compatible with various different compression techniques to achieve this goal, such as GPTQ [25], SparseGPT [23] and AWQ [40]. In our current implementation, we follow the optimal brain surgeon [31, 37] approach and leverage SparseGPT [23] (since it has support for both sparsification and quantization) to compute the optimal $\mathbf{M}$ and $\mathbf{Q}$ by solving the following optimization problem:

$$\arg\min_{\tilde{\Delta}} ||\Delta \cdot \mathbf{X} - \tilde{\Delta} \cdot \mathbf{X}||_2^2 \tag{1}$$

where $\tilde{\Delta}$ is the compressed delta and $\mathbf{X}$ is the input to the layer, which is from the calibration set used for compression.

The major distinction from full model compression is that ΔCompress reconstructs the weight matrix for each layer (Line 6) after compressing the delta and computes the input for next layer (Line 7). This is because the input data (i.e., $\mathbf{X_n}$ in Line 5) is crucial for the compression algorithm. Without re-adding the base weight matrix and reconstructing the weight matrix for each layer, the magnitude of the deltas causes diminishing outputs, leading to vanishing activations in deeper layers. The vanishing activations, as the input to the next layer, will make the compression algorithm fail to capture the input (i.e., X_n Line 5). By extracting the delta and reconstruting the weight for each layer on the fly, ΔCompress ensures proper calibration and maintains high model quality.

Beyond model quality, ΔCompress has two advantages: 1) **Low memory requirement**. As ΔCompress performs layer-wise compression, it only needs a GPU with sufficient memory for a single layer to perform the forward pass. 2) **No need to retrain the model**. Unlike some other compression algorithms [11, 65], ΔCompress does not require further fine-tuning the model to recover the model quality.

5 Serving System Design

We implement DeltaZip in 18K lines of Python and 1K lines of C++/CUDA code. We build the serving engine on top of vLLM [36], HuggingFace Transformers [69], and Sparse Marlin [10]. §5.1 explains how DeltaZip decouples base and delta serving to maximize request batching for base model inference. §5.2 describes how the system parallelizes low-precision delta serving with a custom GPU kernel design. Finally, we describe how we extend model parallelism for delta serving (§5.3) and schedule delta serving requests with continuous batching (§5.4).

5.1 Base and Delta Decoupling

DeltaZip's serving engine always keeps the base model in GPU memory,[1] and swaps compressed deltas on-demand to serve inference requests. The naive approach to serve a fine-tuned model is to load its compressed delta, decompress it, add it to the base model, and then perform inference. However, this approach is inefficient for several reasons: 1) it requires decompressing the delta on the critical path, which adds latency; 2) it does not allow batching requests to different model variants with the same base model; 3) performing inference after adding the delta back to the full-precision base model does not leverage low-precision computation to reduce delta inference latency; and 4) storing decompressed deltas in GPU memory to add back to the base model limits how many deltas can fit concurrently in the GPU.

To improve inference latency and throughput, our first optimization is to **decouple the inference computation of the base model and delta**. Consider a matrix multiplication, which we can decouple as follows by the distributive law:

$$\begin{aligned} \mathbf{Y} &= \mathbf{w}_{\text{fine-tuned}}\mathbf{X} = (\mathbf{w}_{\text{base}} + \Delta)\mathbf{X} \\ &\approx \underbrace{\mathbf{w}_{\text{base}} \cdot \mathbf{X}}_{\text{Batched FP16 matmul}} + \underbrace{\Delta \cdot \mathbf{X}}_{\text{Quantized and sparse matmul}} \end{aligned} \tag{2}$$

In Eq 2, $\mathbf{w}_{\text{base}}\mathbf{X}$ refers to the matmul with the base model, which is shared across all fine-tuned models and we can compute this with a standard `GEMM`. $\Delta\mathbf{X}$ denotes the computation with the delta, which is in a sparse and low-precision format. DeltaZip decouples the computation into two parts and executes the batched base model and low-precision delta matrix multiplications independently and in parallel.

[1] If there are M base models, we divide the GPU cluster into M sets of GPUs, each dedicated to serving a particular base model's variants.

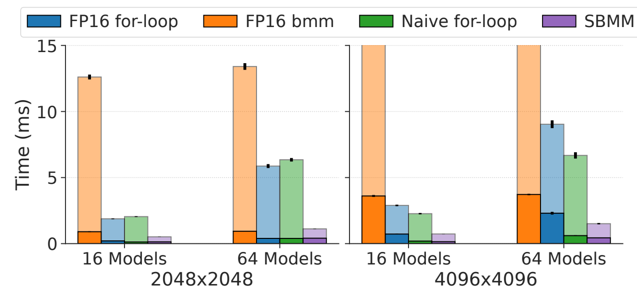

Figure 7. Breakdown of total execution time for different implementations of batched matrix multiplication. The dark part of the bar shows the portion of total execution time spent on computation. Naive for-loop refers to performing low-precision and sparse computation while looping through all models. SBMM refers to our proposed kernel.

Since the distributive law does not hold for non-linear operations, such as activation functions, we decouple the computation at the granularity of linear layers. We merge results from the base model and the delta part after each linear layer to get the output to feed into a non-linear operation. In a transformer block, we serve all linear layers with low-precision, such as the QKV projections $\mathbf{w_q}, \mathbf{w_k}, \mathbf{w_v}$, output projection $\mathbf{w_o}$ and the linear layers in the MLP module.

Decoupling base and delta serving improves GPU utilization and performance in several ways. First, for the base model computation, it enables batching requests for different model variants, as long as they share the same base model. Second, for the delta computation, keeping deltas in low-precision, sparse formats allows us to fit more deltas in GPU memory and reduce swapping. Third, low-precision delta serving also reduces inference latency since token generation in LLM inference is inherently memory-bound (as discussed in §2.1), so the decoding latency is proportional to the GPU memory consumption of the model weights.

5.2 GPU-Efficient Delta Serving

DeltaZip parallelizes the delta computation (ΔX) in a GPU-efficient manner with an approach we call Selective Batched Matrix Multiplication (*SBMM*). For a single batch of requests $\mathbf{X}_i^j$ where i is the request index and j is the delta index, the delta computation can be formalized as: given a batch of requests $\mathbf{X}_1^1, \mathbf{X}_2^2, \ldots, \mathbf{X}_i^j$, compute $\mathbf{Y}_1 = \mathbf{X}_1^1 \cdot \Delta_1, \ldots, \mathbf{Y}_i = \mathbf{X}_i^j \cdot \Delta_j$. We use an index Idx_i to denote the delta index for each request i. Naively, we can loop through different deltas in the batch, find the respective requests and compute the matrix multiplication. However, we observe that this approach is inefficient for two main reasons: 1) it introduces random memory accesses to fetch the inputs and write the outputs

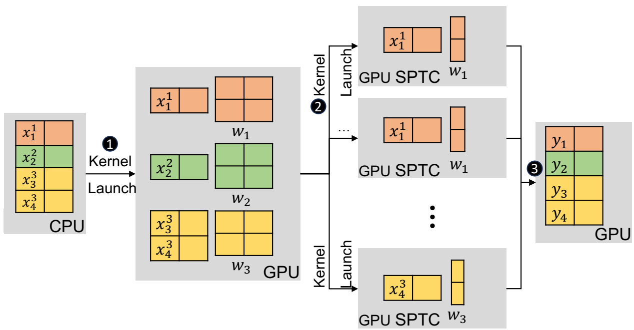

Figure 8. SBMM kernel launch. The kernel computes the output $y_i = \text{matmul}(x_i, w_{Idx_i})$ where the matmul is low precision and sparse. Different colors represent different deltas and their respective requests. SPTC=sparse tensor cores.

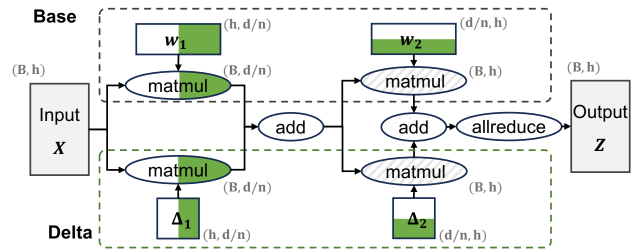

Figure 9. Tensor Parallelism in DeltaZip for $n = 2$ GPUs. B = number of tokens, h = input dimension, d = hidden size. Column-wise and row-wise partitions are illustrated as vertically and horizontally divided boxes, respectively.

to the correct locations, 2) it needs to launch the matrix multiplication kernel multiple times which computes on a small number of requests each time, incurring a high overhead and low GPU utilization. Another option is to use an operator with a batch dimension like `torch.bmm` [22]. However, this requires first stacking the weight matrices for each input into a single matrix, which is not efficient or scalable with delta matrices, as they have large memory footprints. Figure 7 shows the total execution time of different batched matrix multiplication implementations and the portion of time each spends on computation. We find that although the low-precision matmul kernel reduces computation time, the total execution time is still high as it is dominated by kernel launch time and other overheads.

To reduce this overhead and improve GPU utilization, we apply two optimizations. First, before launching a computation on the GPU, DeltaZip's job scheduler **reorders the requests to group requests belonging to the same delta together**. This reduces random data accesses during computation and enables higher batch sizes for delta serving.

Second, we **design a kernel that performs the SBMM operation for multiple deltas in a single kernel launch**. We design SBMM to compute its kernel launch configuration dynamically to balance resources for each delta, since there is often a different number of requests for each delta in the batch. We implement this with CUDA dynamic parallelism [4] on modern GPUs. Specifically, SBMM first launches a kernel that prepares the launch configuration, pointers to the weight, input and output addresses, and other necessary information for each delta, and then launches the actual blocked matmul kernel, which fuses dequantization for each delta and leverages sparse tensor cores on GPUs. The dynamic parallelism feature is revamped in the recent CUDA toolkit [6] and offers substantial performance improvements. Figure 8 illustrates the kernel launch process for 3 deltas and 4 requests where the third delta has two requests. The first kernel is launched from the host and prepares the addresses of the weights, input and output for each delta, and then launches a blocked matmul for each delta in the second step. The actual matmul, in the last step, writes the results to the output. Figure 7 shows that even though the actual computation time is similar, our optimized kernel significantly reduces overhead and improves end-to-end latency.

We design our SBMM kernel to be compatible with popular low-precision and sparse matrix multiplication implementations. Optimizing such computations is an active research area for which many libraries have been developed, such as BitBLAS [67], Marlin [24], and SparseMarlin [10]. Maintaining compatibility with these libraries allows us to leverage the latest hardware and community efforts.

5.3 Model Parallelism for Delta Serving

To serve large models that do not fit into a single GPU, DeltaZip extends Megatron-style [60] tensor parallelism to serve compressed deltas. In Megatron-style tensor parallelism, the model is partitioned column-wise or row-wise across multiple GPUs. DeltaZip adapts this to delta serving by **partitioning the delta in the same way as the base model**. We first illustrate how our partition strategy works in Figure 9 with two linear layers, and then explain how we extend this approach to self-attention layers. Figure 9 assumes that we have two GPUs, a base model with two linear layers $[\mathbf{w}_1, \mathbf{w}_2]$ and a delta $[\Delta_1, \Delta_2]$. In the upper box, we partition the base model $\mathbf{w}_1$ to $[\mathbf{w}_{1,1}, \mathbf{w}_{1,2}]$ by column across two GPUs, and calculate the partial results $Y_{\text{base},i} = X\mathbf{w}_{1,i}$ on each GPU. In the lower box, we partition the delta Δ_1 to $[\Delta_{1,1}, \Delta_{1,2}]$ in the same way, and calculate the partial results $Y_{\text{delta},i} = X\Delta_{1,i}$ on each GPU. The result of the matrix multiplication can be computed on each GPU independently without any synchronization with other GPUs as $Y = [Y_1, Y_2] = [Y_{\text{base},1} + Y_{\text{delta},1}, Y_{\text{base},2} + Y_{\text{delta},2}]$.

We then partition the second linear layer $\mathbf{w}_2, \Delta_2$ with row-parallel across two GPUs as $[\mathbf{w}_{2,1}, \mathbf{w}_{2,2}]^T, [\Delta_{2,1}, \Delta_{2,2}]^T$. Then the output of this layer becomes $Z = [Y_1, Y_2] \cdot [\mathbf{w}_{2,1} + \Delta_{2,1}, \mathbf{w}_{2,2} + \Delta_{2,2}]^T$. To compute this, we first perform $Y_i \cdot \mathbf{w}_{2,i}$ and $Y_i \cdot \Delta_{2,i}$ on each GPU individually, and then sum the results across GPUs with an all-reduce operation as the output.

In the self-attention module, we partition the $\mathbf{q}, \mathbf{k}$ and $\mathbf{v}$ projections as column-wise linear layers and the output projection **o** as row-wise linear layers. We then apply the same strategy as we described above to compute the output of the transformer block.

5.4 Continuous Batching and Scheduling with Delta

Last but not least, DeltaZip optimizes inference by extending continuous batching [76], a standard technique to improve GPU utilization for LLM serving. DeltaZip implements a software scheduler per set of GPU workers that form a tensor parallel serving group. At every iteration, the scheduler picks N models to serve concurrently on a first-come-first-served basis, where N is at most the number of deltas that can fit concurrently in GPU memory across the workers. After selecting the first N deltas, the scheduler searches the request queue for all requests that belong to the selected N models. To maximize batching, it allows requests to *skip the line* if they are for one of the N selected models at the head of the queue. The scheduler delivers the batch of requests to the Model Runner, which starts to load the requested deltas, performs prompt processing if needed, and then batches request decoding for different models.

Tuning N, the number of concurrent deltas. DeltaZip tunes N empirically based on a short trace profiling phase. We explore the impact of varying N. Intuitively, processing more deltas concurrently (larger N) allows for more extensive request batching, enhancing throughput. However, collocating many deltas increases GPU memory pressure, potentially degrading performance if there are already many requests to batch for a particular delta. On the left side of Figure 10, we show how N affects the serving latency when executing a 25 second time interval of a trace with arrival rate 3 and zipf-4.0 model popularity distribution. In this case, DeltaZip would pick $N = 3$ as it achieves optimal performance during profiling. We then run a series of traces under different settings. The right side of Figure 10 shows that $N = 3$ remains optimal or near-optimal across different arrival rates and model popularity distributions. We generally find that offline profiling with a short trace is sufficient to determine the (near-)optimal number of concurrent deltas. Dynamic tuning can also be implemented.

Avoiding starvation. Allowing requests to skip the line is good for batching, but it may cause starvation if requests for the currently selected deltas keep arriving and skipping the line before the system has a chance to swap and serve other deltas. When a request skips the line to be batched with a currently loaded delta, we refer to the original request for that delta at the head of the queue as the *parent request*. To alleviate starvation, DeltaZip uses a preemption strategy: **requests that skip the line are preempted when their parent request finishes**. Preempted requests get reinserted in their original place in the request queue (as if they did not skip the line) and can get scheduled in the next iteration. DeltaZip currently swaps the intermediate state of preempted requests to CPU memory and resumes computation when the request is re-scheduled. Future work involves exploring whether and when recomputing from scratch may be faster than swap-and-resume.

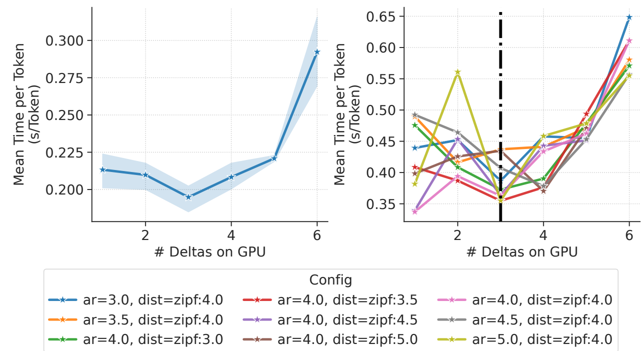

Figure 10. Normalized latency of DeltaZip with varying N, the number of concurrent deltas in GPU memory.

6 Evaluation

6.1 Performance Evaluation Setup

Experiment testbed. We conduct our experiments on a homogeneous GPU cluster. Each node has 2× Intel Xeon 8358P CPUs (128 threads) and 2TB DDR4 memory. We use 4 A800 GPUs per node and GPUs are interconnected to each other by NVLink and NVSwitch [2]. Additionally, the cluster adopts an all-NVMe shared parallel file system, ensuring rapid data access and storage, connected through a 50Gbps RoCE network. All the experiments are conducted on this cluster unless explicitly stated.

Models and downstream tasks. We use the Llama-2 [64] model with 7B, 13B, and 70B parameter and their fine-tuned variants. For 7B and 13B models we use the `Vicuna-v1.5` [15] fine-tuned models since the fine-tuning data is disclosed and can be used for calibration. For 70B model we use the `Llama-2-70B-chat-hf` variant, and we use the fine-tuning dataset from `Vicuna` [15] as a proxy to calibrate the compression. We mainly focus on serving 7B and 13B models in the serving experiments and compress them to 4-bit with 50% sparsity. For these large models, we evaluate the post-compression quality on standard benchmarks in *lm-eval-harness* [27]. We also fine-tune a smaller scale `Pythia-2.8B` model on *natural instructions* tasks and evaluate the quality on the known downstream tasks. Since the accuracy drop is already substantial for SparseGPT and we found that the accuracy drop is more significant with lower precision, we do not evaluate the 2-bit 50% sparsity for SparseGPT.

Workload traces. To evaluate the serving performance, we use the prompts and responses sampled from the LMSys Chatbot Arena [81]. We consider three types of model popularity distribution: 1) Uniform: all models are equally popular. 2) Skewed: model popularity follows a Zipf-α distribution. In other words, the popularity of the i-th model

Base Model	Method	Downstream Tasks↑			Compress Ratio
		T1	T2	T3	
Pythia 2.8B	FP16	73.76	94.23	79.61	1.00×
	SparseGPT (4bit⋆)	67.59	91.99	72.67	3.93×
	DeltaZip (4bit⋆)	73.13	**94.30**	**79.52**	4.75×
	DeltaZip (2bit⋆)	**74.44**	94.22	78.90	**8.36×**
Llama 7B	FP16	80.92	41.65	27.34	1.00×
	SparseGPT (4bit⋆)	67.16	36.48	24.12	5.61×
	AWQ (4bit)	80.86	41.66	27.96	3.64×
	DeltaZip (4bit⋆)	81.41	41.72	27.50	5.39×
	DeltaZip (2bit⋆)	**81.56**	**41.91**	**28.26**	**10.36×**
Llama 13B	FP16	85.29	43.00	27.04	1.00×
	SparseGPT (4bit⋆)	79.88	35.01	23.20	5.91×
	AWQ (4bit)	84.80	**43.37**	**27.80**	3.82×
	DeltaZip (4bit⋆)	**85.11**	42.48	27.04	5.91×
	DeltaZip (2bit⋆)	84.95	42.54	27.65	**11.83×**
Llama 70B	FP16	86.73	44.25	33.18	1.00×
	SparseGPT (4bit⋆)	85.87	37.06	27.80	6.11×
	AWQ (4bit)	86.36	44.04	31.80	3.72×
	DeltaZip (4bit⋆)	**87.28**	**44.18**	32.87	5.84×
	DeltaZip (2bit⋆)	86.67	43.47	**33.49**	**13.96×**

Table 1. Model quality of DeltaZip vs. uncompressed (FP16) vs. SparseGPT [23] and AWQ [40]. For Pythia, T1, T2, T3 show the accuracy on 3 downstream tasks (Amazon Review Classification, Synthetic Palindrome Numbers, Yes/No Question) from *natural instructions* [45]. For other models, T1, T2, T3 show accuracy on 3 standard benchmarks (BoolQA, TruthfulQA, LogiQA) in *lm-eval-harness* [27]. ⋆ indicates 50% structured pruning and quantization.

is proportional to $1/i^{\alpha}$. We set $\alpha = 1.5$ for the skewed distribution. 3) Azure trace: since there is no publicly available traces for multi-variant LLM serving available, following previous work [39], we use the Azure serverless function traces [58, 79] as a proxy. Note the traffic in this trace is highly bursty and the model distribution is highly skewed. Except for the small-scale timeline analysis in Figure 16 (which we show for illustration purposes) and ablation studies, we run the traces for 5 minutes under different arrival rates and model distribution. Unless otherwise stated, we assume there are 32 model variants that need to be served.

Metrics. We use downstream accuracy to evaluate the post-compression quality and three metrics for serving performance: 1) average end-to-end (E2E) latency, 2) average time to first token (TTFT), which is an important indicator of the system's responsiveness, 3) throughput and 4) SLO attainment (i.e., percentage of requests served within a given SLO requirement) for TTFT and E2E Latency.

Baselines. For compression quality, we compare DeltaZip with SparseGPT [23] which incorporates both quantization and sparsification, as well as AWQ [40] which is a state-of-the-art quantization algorithm. For serving performance,

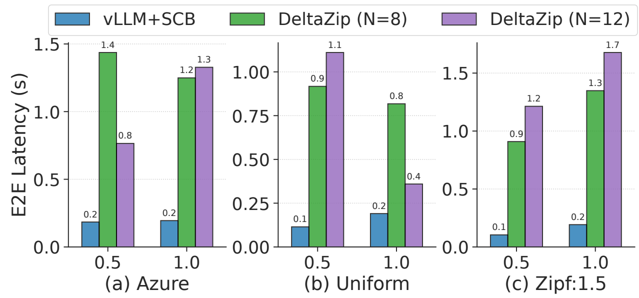

Figure 11. Throughput of different serving systems with varying poisson arrival rate $\lambda \in \{0.5, 1.0\}$ and distribution $\mathcal{D} \in \{\text{azure, uniform, zipf-1.5}\}$ for 13B model.

since there is no existing system that can serve multiple full-parameter fine-tuned models, we implement a simple baseline based on vLLM that supports 1) Swapping models in and out of GPU memory, 2) Continuous batching of different full-precision models by looping through models in a batch and 3) Batching available requests for the same model. We refer to this baseline as vLLM-SCB. We use the maximum number of models that can fit in the GPU memory for the baseline serving system. We use a tensor parallelism degree of 4 for both DeltaZip and the baseline system unless otherwise stated. We do not show a comparison to ServerlessLLM [26] as it treats each model as a black box — meaning it does not batch requests from models derived from the same base — and hence is not competitive for the scenarios we explore.

6.2 Post-Compression Model Quality

We first study how inference accuracy is impacted by applying quantization and sparsification on model deltas compared to directly on FMT model weights. As shown in Table 1, DeltaZip achieves up to 13× compression ratio with comparable accuracy to the original FP16 FMT models. On smaller models, DeltaZip achieves 8× to 11× compression ratio with comparable accuracy. In contrast, SparseGPT — which applies similar techniques directly to the FMT model weights — has a substantial drop in accuracy for all models. DeltaZip also achieves comparable accuracy compared with the state-of-the-art quantization algorithm AWQ [40] but with a much higher compression ratio. We also observe that in some cases, the accuracy of the compressed model is even higher than the full-precision model. This effect is also noticeable in other works [25, 30, 40] and is likely due to the regularizaiton effects of compression. These results show that ΔCompress can aggressively compress model deltas while retaining higher accuracy than the traditional approach of compressing model weights directly. In §6.4, we also show that compressed delta serving achieves higher accuracy on complex tasks than LoRA serving.

6.3 End-to-end Serving Performance

Serving throughput. Figure 11 shows the throughput of different serving systems with varying arrival rates and

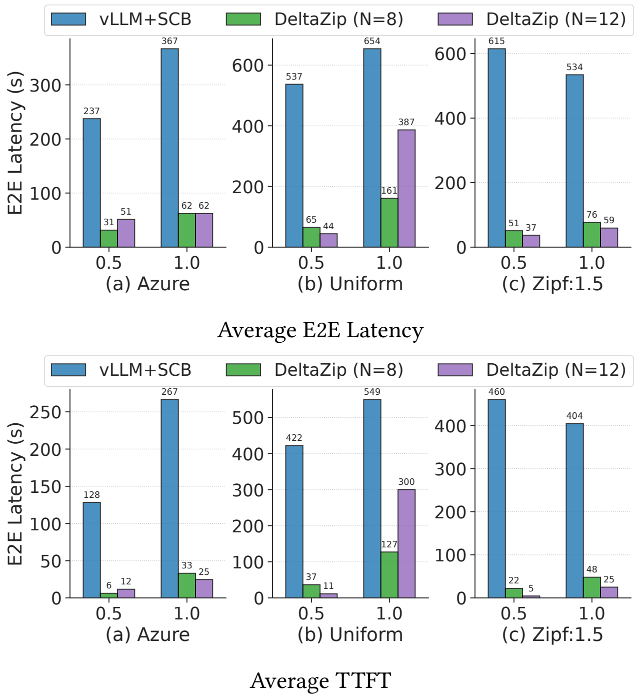

Figure 12. Average latency of different serving systems with varying arrival rate and distribution for 13B model.

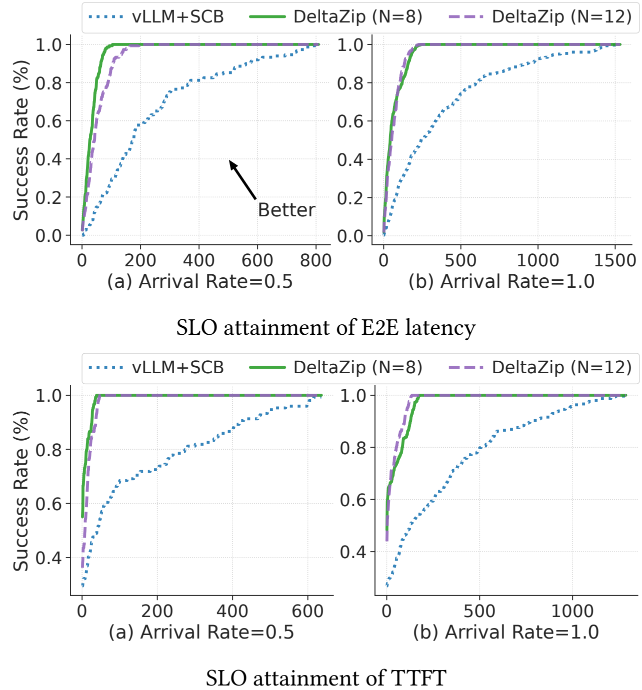

Figure 13. SLO attainment of different serving systems with varying arrival rate and azure model distribution for 13B model. X-axis denotes different SLO requirement in seconds.

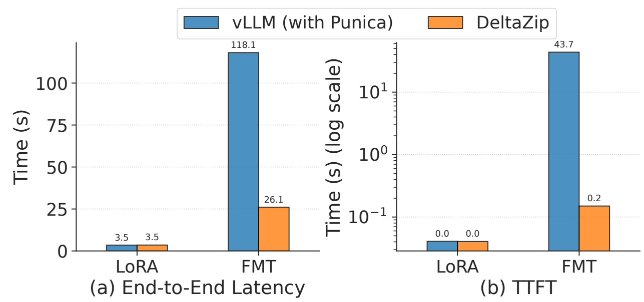

Figure 14. E2E latency and TTFT of DeltaZip and LoRA serving system when serving LoRA and FMT models.

model distributions. We observe a 2× to 12× improvement in throughput compared to the baseline system. The improvement is more pronounced when the arrival rate is relatively light and more skewed. When the arrival rate is high and the model distribution is uniform, the improvement is less significant, and we find this is mainly due to the high cost in prompt processing when the models are more uniform. Since our techniques cannot reduce the prompt processing time, and when more deltas are batched together, requests in this batch have to wait for the prompt processing of the slowest request in the batch, leading to a lower throughput.

Average E2E latency and TTFT. Figure 12 shows that DeltaZip acheives a 1.6× to 16× improvement in average E2E latency and even higher improvement in TTFT. The high improvement in TTFT is due to DeltaZip's ability to batch more requests from different models concurrently, thus reducing queuing time. As in our throughput experiments, the improvement under uniform distribution with high load is less significant. We also observe that the maximum number of concurrent deltas being served has a substantial impact on the performance.

SLO Attainment of TTFT and E2E Latency. Figure 13 shows the SLO attainment of E2E latency and TTFT with the Azure trace under varying Poisson arrival rate. We observe that DeltaZip achieves a higher SLO attainment compared to the baseline system.

6.4 Delta versus LoRA Serving Approaches

Complementary to LoRA serving systems, DeltaZip extends their serving optimizations to also efficiently serve FMT models. DeltaZip inherits the ability to serve LoRA adapters from vLLM [36], which supports Punica-based [12] LoRA serving. Figure 14 shows an example of using DeltaZip

Base Model	**Task**	**Accuracy↑**		
		FMT	LoRA	ΔCompress
Llama-7B	Math (GSM8K)	34.79	29.49	34.95
Pythia-2.8B	Amazon Review	73.76	50.92	73.13
	BoolQ Yes/No	79.61	63.76	79.52
	NLI Classification	73.23	52.15	70.74
OpenLlama 3B	Amazon Review	76.07	55.07	77.36
	BoolQ Yes/No	83.38	65.69	83.29
	NLI Classification	80.07	63.46	79.75

Table 2. Model quality of FMT vs. LoRA vs. ΔCompress.

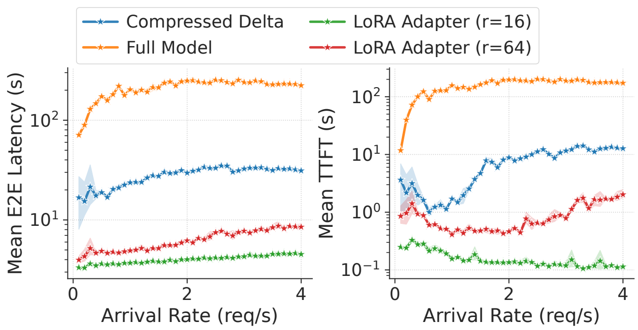

Figure 15. End-to-end latency and TTFT of DeltaZip and LoRA serving system with varying arrival rate.

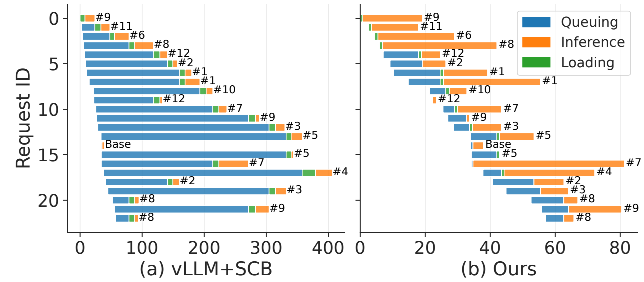

Figure 16. Serving latency breakdown. The # on the right indicates the model ID. The x-axis is the time in seconds.

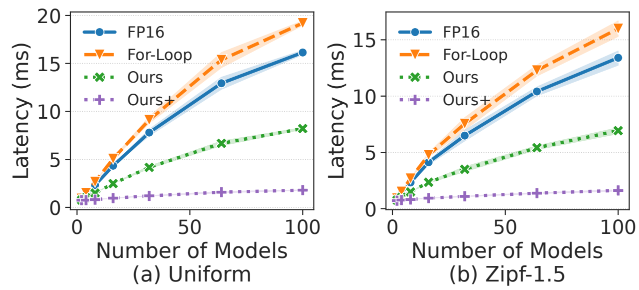

Figure 17. Microbenchmark of the SBMM kernel vs. the baseline implementation on a single GPU with varying number of models given a fixed number of requests. "Ours" refers to the implementation with reduced random memory access only and "Ours+" refers to the kernel implementation we proposed in §5.1. For FP16, we do not perform the decoupling as it will only introduce additional overhead.

to serve LoRA adapters on one GPU node and FMT model variants on another GPU node. For LoRA serving, DeltaZip achieves similar performance as vLLM with Punica kernels and for FMT serving, DeltaZip significantly outperforms the baseline due to its delta compression approach.

Given that DeltaZip supports both LoRA and FMT model serving, a natural question that arises for users is when to use which type of model fine-tuning approach. While a detailed study is outside the scope of this paper and prior works have already compared FMT and LoRA accuracy [5, 9], we conduct a brief analysis in which we apply delta compression for FMT serving. Table 2 compares the accuracy of uncompressed FMT models, LoRA adapters and ΔCompress compressed FMT models. We conduct extensive hyper-parameter tuning for LoRA adapters, using scripts from Anyscale [5]. We observe that even on complicated tasks (e.g., Math) where LoRA adapters cannot achieve similar accuracy as FMT models, ΔCompress can maintaining high accuracy while compressing the FMT models.

We also compare the inference latency and TTFT of LoRA adapter, compressed delta FMT and baseline full model FMT serving with varying arrival rates. Figure 15 shows that compressed deltas and LoRA adapters are much more efficient to serve than the vLLM+SCB baseline approach for FMT model serving, which swaps full models. Serving LoRA adapters is still more efficient than serving compressed deltas, mainly due to the more compact size and smaller memory footprint of LoRA adapters. In conclusion, **users should choose between LoRA and compressed delta FMT serving based on the trade-off between accuracy and serving performance**: LoRA is more suitable for tasks where accuracy is not the primary concern or for simpler tasks where LoRA can achieve comparable accuracy to full model tuning. In contrast, compressed delta FMT serving is more suitable for tasks where accuracy is critical. DeltaZip improves the serving efficiency for FMT serving with the ΔCompress delta compression approach and its optimized delta serving system design and implementation.

6.5 Microbenchmarks and Ablation Study

Latency breakdown. We perform a smaller-scale experiment to visualize the latency breakdown of different serving stages (queuing, loading, and inference). We synthesize a trace with 12 models, arrival rate of 0.5 requests per second for 60 seconds and run on two RTX 3090 GPUs. Figure 16 compares DeltaZip with the baseline. The vLLM+SCB system has two main issues: 1) the loading time is substantial, as it needs to load the entire model from the disk to GPU, and 2) queuing time dominates, due to the lack of batching. Even though it batches requests for the same model (e.g., for the model #2), it cannot batch requests from many different models due to GPU memory capacity limits. In contrast, DeltaZip alleviates these two issues by 1) loading only the compressed deltas, which can be 5× to 10× smaller than the full model, and 2) batching requests from different models, which significantly reduces queuing delays.

SBMM kernel. We evaluate the performance of the SBMM kernel described in §5.1. Figure 17 shows the performance of the SBMM kernel compared to the baseline implementations as we scale the number of models. We observe that the naive for-loop approach does not bring much performance

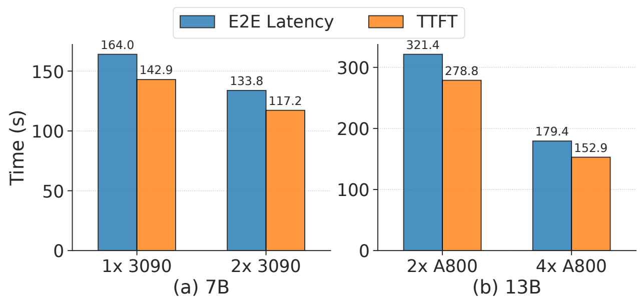

Figure 18. End-to-end latency and time to first token (TTFT) of DeltaZip with varying number of GPUs.

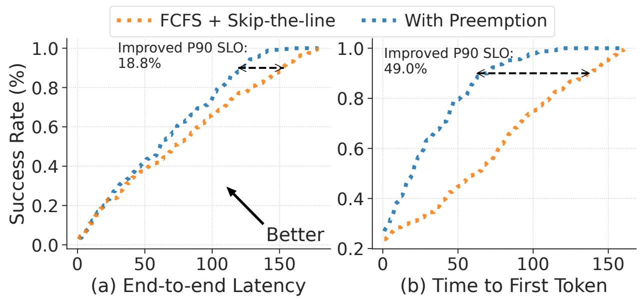

Figure 19. E2E Latency and TTFT of DeltaZip with and without preemption.

improvement and is comparable with the half-precision implementation. However, with reduced random memory access, we observe a 2× speedup compared to the baseline implementation, and the proposed kernel further improves the performance by another 2× to 3×. In addition, we also observe that given the same number of requests, our kernel scales well with different number of models.

Model Parallelism. We also conduct an experiment with varying numbers of GPUs under different settings to evaluate the model parallelism of DeltaZip. Figure 18 shows the end-to-end latency and TTFT of DeltaZip with varying number of GPUs. We observe that the latency decreases with the number of GPUs, particularly on A800 platform. This is because the inter-GPU communication is faster on A800 platform compared to the RTX 3090 platform. This observation leaves a direction for future work to optimize and tune the tensor parallelism degree of DeltaZip.

Starvation Handling. Next we evaluate the effectiveness of the preemption mechanism as described in §5.4. Figure 19 shows the E2E latency and TTFT with and without preemption. We observe that the preemption mechanism effectively reduces the latency, particularly the time-to-first-token as it allows more requests to start earlier.

7 Related Work

ML Model Serving Systems. Optimizing ML serving is an active area of research [16, 29, 57]. With the increasing popularity of LLMs, there has been a surge of LLM serving systems proposing optimizations such as GPU kernel implementations [17, 18, 38, 75], advanced and fine-grained batching [21, 76], memory management [36], and parallelism [39, 60]. However, most of these works do not consider the multi-variant serving scenario, for which DeltaZip is designed. For multi-model LLM serving, MuxServe [20] explores spatial-temporal multiplexing for LLM serving to improve GPU utilization. ServerlessLLM [26] studies the feasibility of serving LLMs in a serverless environment and proposes fast checkpoint loading and locality-driven live migration. However, these works treat the models as a black box and do not consider their lineage, and are hence orthogonal to DeltaZip. Punica and S-LoRA [13, 59] are the most relevant to DeltaZip, as they also target multi-variant model serving scenarios, but they focus on serving LoRA adapters and do not consider serving full fine-tuned models. DeltaZip is complementary to these works.

Post-Training Compression. Another line of work focuses on reducing the memory footprint of large models by compressing the model weights in a lossy manner. Beyond the OBS framework [31] and its improvements, such as GPTQ [25] and SparseGPT [23], many compression algorithms [40, 66, 72, 74] have been proposed to reduce the model size. Such techniques can also be applied to model deltas by adapting the DeltaZip compression pipeline. There is also concurrent work on model delta compression. DeltaZip [73] and GPT-Zip [34] proposes quantization and unstructured sparsity, BitDelta [41] proposes extreme quantization and DARE [77] proposes unstructured sparsity. Compared to these works, DeltaZip optimizes more for **hardware-efficiency** by combining quantization with *structured sparsity*. To the best of our knowledge, DeltaZip is also the first serving system to support both LoRA and compressed delta FMT model serving.

8 Discussion

Limitations. While DeltaZip's decoupled computation improves throughput and reduces latency when concurrently serving many models, decoupled inference still has higher unloaded latency than serving a FMT model directly in GPU memory. Hence, when there are only a handful of models to serve and they fit in GPU memory, DeltaZip may not be suitable. In addition, the co-serving of LoRA and FMT models in DeltaZip is at a coarse granularity, where LoRA and FMT models are served in two separate sets of GPUs and must be in separate batches. We plan to explore the possibility of serving LoRA and FMT models in the same batch as future work.

Supporting PEFT approaches beyond LoRA. DeltaZip's decoupled computation architecture is general and can be used beyond LoRA, for other PEFT methods being proposed to improve accuracy. For example, GaLore [80] only uses low-rank on the gradient but results in full-rank weight updates. RoSA [46] introduces sparse adapters in addition to low-rank adapters. Due to the lack of support for full-rank weight updates, these methods cannot be served by existing LoRA-based systems. We plan to extend DeltaZip to add support for emerging PEFT methods.

9 Conclusion

To conclude, DeltaZip efficiently serves a variety of fine-tuned model variants, whether they are fine-tuned through parameter-efficient or full-model tuning techniques. For efficient serving of full-model-tuned models, DeltaZip leverages a key insight: fine-tuning typically results in small perturbations, allowing model deltas to be highly compressible. DeltaZip co-designs the serving system with the compression algorithm and achieves 10× compression ratio, improves serving throughput by 2× to 12× and maintains high model quality comparable to FP16 models. Our system will be open-source upon publication.

Acknowledgments

We would like to thank Berivan Isik, Hermann Kumbong, Wanyi Ning, Ce Zhang, Sanmi Koyejo for their contribution and insightful feedback during the early stage of the project.

References

[1] Numbers every llm developer should know. https://github.com/ray-project/llm-numbers, 2023.
[2] Nvlink and nvswitch. https://www.nvidia.com/en-us/data-center/nvlink/, 2023.
[3] Meta llama 3. https://github.com/meta-llama/llama3, 2024.
[4] Adinets Andy. CUDA Dynamic Parallelism API and Principles, May 2014.
[5] Anyscale. Fine-Tuning LLMs: LoRA or Full-Parameter? An in-depth Analysis with Llama 2, November 2023.
[6] Rob Armstrong, Rob Nertney, Arthy Sundaram, Matthew Nicely, Robert Jensen, and Fred Oh. Cuda toolkit 12.0 released for general availability. https://developer.nvidia.com/blog/cuda-toolkit-12-0-released-for-general-availability/, 2022.
[7] b mc2. sql-create-context dataset, 2023.
[8] Hongxiao Bai and Yun Li. Structured Sparsity in the NVIDIA Ampere Architecture and Applications in Search Engines, July 2023.
[9] Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less, 2024.
[10] Roberto L. Castro and Dan Alistarh. Sparse-marlin: Boosting 4-bit inference kernels with 2:4 sparsity. https://github.com/IST-DASLab/Sparse-Marlin, 2024.
[11] Yuji Chai, John Gkountouras, Glenn G. Ko, David Brooks, and Gu-Yeon Wei. INT2.1: Towards Fine-Tunable Quantized Large Language Models with Error Correction through Low-Rank Adaptation, June 2023. arXiv:2306.08162 [cs].
[12] Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, and Arvind Krishnamurthy. Punica: Multi-Tenant LoRA Serving, October 2023. arXiv:2310.18547 [cs].
[13] Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, and Arvind Krishnamurthy. Punica: Multi-tenant lora serving. *Proceedings of Machine Learning and Systems*, 6:1–13, 2024.
[14] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code, July 2021. arXiv:2107.03374 [cs].
[15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
[16] Daniel Crankshaw, Xin Wang, Guilio Zhou, Michael J. Franklin, Joseph E. Gonzalez, and Ion Stoica. Clipper: A {Low-Latency} Online Prediction Serving System. pages 613–627, 2017.
[17] Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, July 2023. arXiv:2307.08691 [cs].
[18] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, June 2022. arXiv:2205.14135 [cs].
[19] Thomas Dohmke. Github copilot x: The ai-powered developer experience, March 2023.
[20] Jiangfei Duan, Runyu Lu, Haojie Duanmu, Xiuhong Li, Xingcheng Zhang, Dahua Lin, Ion Stoica, and Hao Zhang. MuxServe: Flexible Multiplexing for Efficient Multiple LLM Serving, April 2024. arXiv:2404.02015 [cs].
[21] Jiarui Fang, Yang Yu, Chengduo Zhao, and Jie Zhou. TurboTransformers: An Efficient GPU Serving System For Transformer Models, February 2021. arXiv:2010.05680 [cs].
[22] PyTorch Foundation. torch.bmm — PyTorch 2.3 documentation.
[23] Elias Frantar and Dan Alistarh. SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot, March 2023. arXiv:2301.00774 [cs].
[24] Elias Frantar and Dan Alistarh. Marlin: a fast 4-bit inference kernel for medium batchsizes. https://github.com/IST-DASLab/marlin, 2024.
[25] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers, March 2023. arXiv:2210.17323 [cs].
[26] Yao Fu, Leyang Xue, Yeqi Huang, Andrei-Octavian Brabete, Dmitrii Ustiugov, Yuvraj Patel, and Luo Mai. Serverlessllm: Locality-enhanced serverless inference for large language models. *arXiv preprint arXiv:2401.14351*, 2024.
[27] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023.
[28] GitHub. GitHub Copilot · Your AI pair programmer.
[29] Arpan Gujarati, Reza Karimi, Safya Alzayat, Wei Hao, Antoine Kaufmann, Ymir Vigfusson, and Jonathan Mace. Serving DNNs like Clockwork: Performance Predictability from the Bottom Up. pages 443–462, 2020.
[30] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
[31] B. Hassibi, D.G. Stork, and G.J. Wolff. Optimal Brain Surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pages 293–299 vol.1, March 1993.

[32] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. arXiv:2106.09685 [cs].

[33] Itay Hubara, Brian Chmiel, Moshe Island, Ron Banner, Joseph Naor, and Daniel Soudry. Accelerated Sparse Neural Training: A Provable and Efficient Method to Find N:M Transposable Masks. In *Advances in Neural Information Processing Systems*, volume 34, pages 21099–21111. Curran Associates, Inc., 2021.

[34] Berivan Isik, Hermann Kumbong, Wanyi Ning, Xiaozhe Yao, Sanmi Koyejo, and Ce Zhang. GPT-Zip: Deep Compression of Finetuned Large Language Models. July 2023.

[35] Juraj Juraska, Kevin K. Bowden, and Marilyn Walker. ViGGO: A Video Game Corpus for Data-To-Text Generation in Open-Domain Conversation, October 2019. arXiv:1910.12129 [cs].

[36] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, New York, NY, USA, October 2023. Association for Computing Machinery.

[37] Yann LeCun, John Denker, and Sara Solla. Optimal Brain Damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

[38] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers, 2022.

[39] Zhuohan Li, Lianmin Zheng, Yinmin Zhong, Vincent Liu, Ying Sheng, Xin Jin, Yanping Huang, Zhifeng Chen, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. {AlpaServe}: Statistical Multiplexing with Model Parallelism for Deep Learning Serving. pages 663–679, 2023.

[40] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, MLSys '24, 2024.

[41] James Liu, Guangxuan Xiao, Kai Li, Jason D Lee, Song Han, Tri Dao, and Tianle Cai. Bitdelta: Your fine-tune may only be worth one bit. *arXiv preprint arXiv:2402.10193*, 2024.

[42] Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Do Emergent Abilities Exist in Quantized Large Language Models: An Empirical Study, July 2023. arXiv:2307.08072 [cs].

[43] Tiedong Liu and Bryan Kian Hsiang Low. Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks, May 2023. arXiv:2305.14201 [cs].

[44] madiepev. Fine-tune a foundation model from the model catalog in Azure Machine Learning - Training.

[45] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-Task Generalization via Natural Language Crowdsourcing Instructions, March 2022. arXiv:2104.08773 [cs].

[46] Mahdi Nikdan, Soroush Tabesh, Elvir Crnčević, and Dan Alistarh. RoSA: Accurate Parameter-Efficient Fine-Tuning via Robust Adaptation, June 2024. arXiv:2401.04679 [cs].

[47] NVIDIA. NVCOMP, March 2021.

[48] OpenAI. GPT-3.5 Turbo fine-tuning and API updates.

[49] OpenAI. Introducing ChatGPT.

[50] OpenAI. OpenAI Pricing.

[51] OpenAI. GPT-4 Technical Report, March 2023. arXiv:2303.08774 [cs].

[52] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. arXiv:2203.02155 [cs].

[53] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative LLM inference using phase splitting, May 2024. arXiv:2311.18677 [cs].

[54] Google Cloud Platform. Vertex AI.

[55] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, December 2023. arXiv:2305.18290 [cs].

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, September 2023. arXiv:1910.10683 [cs, stat].

[57] Francisco Romero, Qian Li, Neeraja J. Yadwadkar, and Christos Kozyrakis. {INFaaS}: Automated Model-less Inference Serving. pages 397–411, 2021.

[58] Mohammad Shahrad, Rodrigo Fonseca, Inigo Goiri, Gohar Chaudhry, Paul Batum, Jason Cooke, Eduardo Laureano, Colby Tresness, Mark Russinovich, and Ricardo Bianchini. Serverless in the Wild: Characterizing and Optimizing the Serverless Workload at a Large Cloud Provider. pages 205–218, 2020.

[59] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. S-LoRA: Serving Thousands of Concurrent LoRA Adapters, November 2023. arXiv:2311.03285 [cs].

[60] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, March 2020. arXiv:1909.08053 [cs].

[61] Foteini Strati, Sara Mcallister, Amar Phanishayee, Jakub Tarnawski, and Ana Klimovic. D\'ej\'aVu: KV-cache Streaming for Fast, Fault-tolerant Generative LLM Serving, March 2024. arXiv:2403.01876 [cs].

[62] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[63] Gemini Team. Gemini: A Family of Highly Capable Multimodal Models, May 2024. arXiv:2312.11805 [cs].

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. arXiv:2302.13971 [cs] version: 1.

[65] Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *arXiv preprint arXiv:2402.04396*, 2024.

[66] Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even Better LLM Quantization with Hadamard Incoherence and Lattice Codebooks, June 2024. arXiv:2402.04396 [cs].

[67] Lei Wang, Lingxiao Ma, Shijie Cao, Quanlu Zhang, Jilong Xue, Yining Shi, Ningxin Zheng, Ziming Miao, Fan Yang, Ting Cao, Yuqing Yang, and Mao Yang. Ladder: Enabling Efficient Low-Precision Deep Learning Computing through Hardware-aware Tensor Transformation. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, 2024.

[68] Panqi (William) Wei. PanQiWei/AutoGPTQ, December 2023. original-date: 2023-04-13T02:18:11Z.

[69] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf,

Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[70] Bingyang Wu, Ruidong Zhu, Zili Zhang, Peng Sun, Xuanzhe Liu, and Xin 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, 2024.

[71] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[72] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models, June 2023. arXiv:2211.10438 [cs].

[73] Xiaozhe Yao and Ana Klimovic. Deltazip: Multi-tenant language model serving via delta compression. *arXiv preprint arXiv:2312.05215*, 2023.

[74] Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers, June 2022. arXiv:2206.01861 [cs].

[75] Zihao Ye, Lequn Chen, Ruihang Lai, Yilong Zhao, Size Zheng, Junru Shao, Bohan Hou, Hongyi Jin, Yifei Zuo, Liangsheng Yin, Tianqi Chen, and Luis Ceze. Accelerating self-attentions for llm serving with flashinfer, February 2024.

[76] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. Orca: A Distributed Serving System for {Transformer-Based} Generative Models. pages 521–538, 2022.

[77] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

[78] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method, February 2024. arXiv:2402.17193 [cs].

[79] Yanqi Zhang, Inigo Goiri, Gohar Irfan Chaudhry, Rodrigo Fonseca, Sameh Elnikety, Christina Delimitrou, and Ricardo Bianchini. Faster and Cheaper Serverless Computing on Harvested Resources. In *Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles*, SOSP '21, pages 724–739, New York, NY, USA, October 2021. Association for Computing Machinery.

[80] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

[81] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, July 2023. arXiv:2306.05685 [cs].

[82] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving, June 2024. arXiv:2401.09670 [cs].

[83] Aojun Zhou, Yukun Ma, Junnan Zhu, Jianbo Liu, Zhijie Zhang, Kun Yuan, Wenxiu Sun, and Hongsheng Li. Learning N:M Fine-grained Structured Sparse Neural Networks From Scratch, April 2021. arXiv:2102.04010 [cs].

[84] Xuanhe Zhou, Guoliang Li, and Zhiyuan Liu. LLM As DBA, August 2023. arXiv:2308.05481 [cs].

[85] Zhe Zhou, Xuechao Wei, Jiejing Zhang, and Guangyu Sun. {PetS}: A Unified Framework for {Parameter-Efficient} Transformers Serving. pages 489–504, 2022.