Transit-Layer Tensor Representation for Serverless Inference

Ben Mechels
University of Minnesota

Shu Shi
University of Chicago

Wenyi Wang
University of Chicago

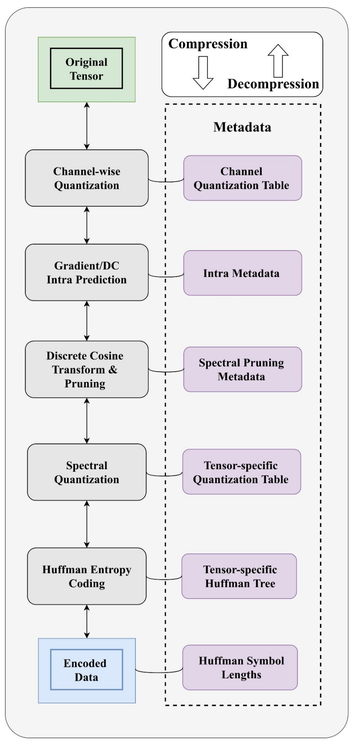
Kyle Chard
University of Chicago

1 Introduction

- Serverless inference frameworks schedule LLM tensor loading and swapping in many-user, many-model environments
- Moving models is slow. Existing techniques have reached a ceiling in scheduling and loading optimizations, and are now bottlenecked by hardware transfer speed
- We introduce `Condense`, a new approach to compress tensor checkpoint representations while in transit to accelerate the loading and swapping latencies of serverless frameworks
- We design a novel compressor pipeline and integrate it into the ServerlessLLM runtime then evaluate the improvements across LLaMa models

2 Pipeline Design

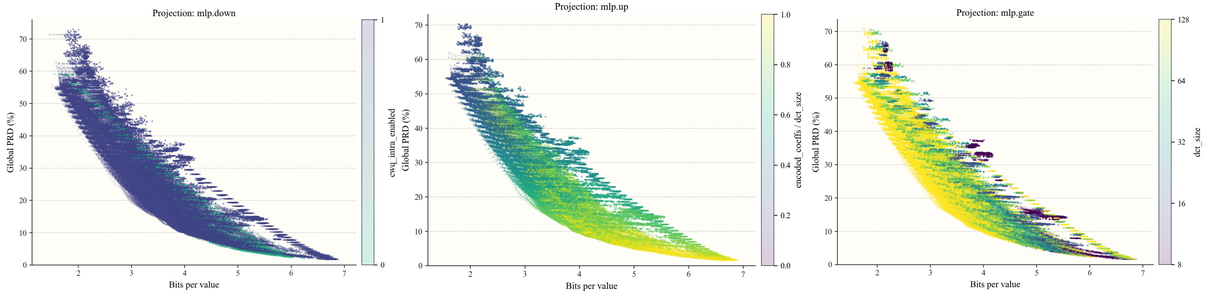
Design Objective: High-throughput block independent processing with emphasis on metadata extraction



Design Features:

- ➔ Parameterization done per weight projection
- ➔ Explicitly-stored and globally set prediction policy
- ➔ Adaptive spectral pruning by least-*k* selection per transform window
- ➔ Outlier-mitigating quantization tables
- ➔ Low overhead, fast Symlen-style Huffman coding
- ➔ Full GPU support for data path with only structure parsing on CPU

3 Pipeline Validation

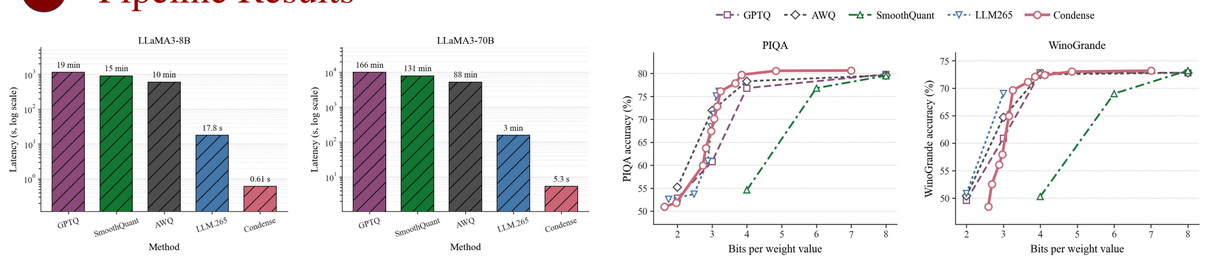


Best performing rate-distortion points all require prediction

Pruning ratio is effective method of navigating rate-distortion curve

Large 1D transforms capture tensor context better than other methods

4 Pipeline Results



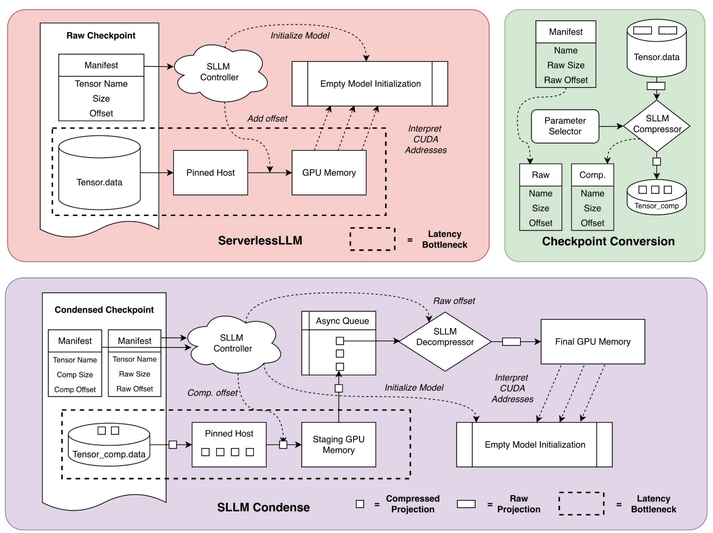
`Condense` significantly improves runtime over state-of-the-art quantization and compression methods

Even on fixed-bitrate mode, `Condense` matches downstream task accuracies of existing techniques

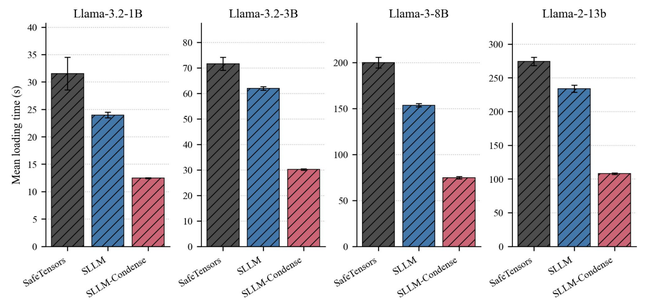
5 Framework Design

Design Objective: Reduce transfer latency with minimal post-processing overhead

- ➔ `Condense` uses a compact tensor representation to alleviate disk read bottleneck by reducing the size of transferred data
- ➔ `Condense` utilizes async. queuing of weight projections to reduce tail operations post-copy to only milliseconds of runtime
- ➔ Due to stage overlap, `Condense` allocates working-space memory for the decompressor binding



6 Framework Results



- ➔ `Condense` provides a consistent speedup over ServerlessLLM in cold-start checkpoint loading regardless of model size
- ➔ Latency reduction is roughly proportional to ratio of compressed checkpoint size to raw tensor size

7 Future Work

Pipeline

1. Sparse pruning trees to achieve sub 2-bit compression
2. Fused kernels to reduce DRAM bottleneck

Framework

1. Modify inference engine to support *Condense* format
2. On-the-fly KV cache compression for live migration

8 Conclusion

- ➔ Compressed transit layer representations improve latency with negligible post processing overhead
- ➔ Designing pipeline stages with emphasis on throughput provides comparable rate-distortion performance at a fraction of the runtime
- ➔ We provide a novel compressor design alongside a new tensor compression application and demonstrate that both can be co-designed

References

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. Proceedings of the 40th International Conference on Machine Learning.
Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. In The Eleventh International Conference on Learning Representations
Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. In *Proceedings of Machine Learning and Systems* 6 (MLSys 2024).
Ceyu Xu, Yongji Wu, Xinyu Yang, Beidi Chen, Matthew Lentz, Danyang Zhuo, and Lisa Wu Wills. 2025. LLM.265: Video Codecs are Secretly Tensor Codecs. In *Proceedings of the 58th IEEE/ACM International Symposium on Microarchitecture* (MICRO 2025). ACM, 445-460.
Yao Fu, Leyang Xue, Yeqi Huang, Andrei-Octavian Brabete, Dmitrii Ustiugov, Yuvraj Patel, and Luo Mai. 2024. ServerlessLLM: Low-Latency Serverless Inference for Large Language Models. In *Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation* (OSDI 2024). USENIX Association, 135-153.