Breaking the Resource Monopoly from Industries: Sustainable and Reliable LLM Serving By Recycling Outdated and Resource-Constrained GPUs

Tianlong Chen

Department of Computer Science
University of North Carolina at Chapel Hill
tianlong@cs.unc.edu

Abstract

In recent years, Large Language Model (LLM) agents, exemplified by models like ChatGPT, and PaLM, have showcased remarkable prowess in various tasks, owing to their vast number of parameters and emergent in-context learning capabilities. People expect the wide usage of LLM serving at-edge hardware, personal devices, and organization/enterprise IT infrastructures to revolutionize global access to information, communication, automation, and creativity. However, due to the extreme large-scale LLM parameters (LLaMA 3.1 contains 405 billion of 2 or 4 bytes floating point numbers), the LLM serving is facing significant sustainability pressure due to its requirements on the latest high-embodied carbon hardware (*e.g.*, GPUs, HBMs, memory, storage, and network hardware) and the high operational carbon emissions, leading to a significant and alarming increase in carbon emissions and a high barrier to their widespread deployments and practical applications in various scenarios. Companies, organizations, and institutes usually have the complete general-purpose IT infrastructure, which consists of a large amount of computing, memory, storage, and network hardware. Although these general-purpose IT infrastructures are far more than enough for existing application executions, deploying and executing the LLM for a broad spectrum of serving platforms can be challenging and difficult due to resource limitations. Purchasing the latest hardware including GPUs (*e.g.*, Nvidia H100 or H200) will lead to considerable issues including 1) serious embodied carbon emissions during the new hardware production, 2) no explicitly lower operational carbon emissions with essential modeling and optimizations, 3) high economic and financial pressures, and 4) potentially tremendous existing hardware resource wasting. Therefore, **it is a trend and becomes a must to explore how to use the existing hardware, especially outdated hardware, to collectively improve both environmental sustainability, efficiency, and reliability for LLM serving**. A few pioneering examples include Microsoft's Project Natick, Google's TPU Pod Optimization, Alibaba's Cloud Server Repurposing, and Facebook's Network Hardware Reuse. This talk will particularly emphasize *modularized LLM architecture*, *in-storage sustainable computing*, and *reliable serving against software and hardware attacks*.

Structure of the Talk

▷ **(Part 1 — 10 mins) Modularize, Allocate, and Compress: Enable Sustainable LLM Serving on Resource-Constrained GPUs.** Initiating our journey. I will briefly introduce the challenges of LLM serving on outdated and resource-constrained GPUs. To address these pain points, I will review recent advances in modularizing, allocating, and compressing LLMs for sustainable serving. I will first explain how to convert vanilla LLMs into their modularized counterparts without sacrificing the serving quality [1]. Then, I will cover solutions to advance modular network allocations by tackling unbalanced routing, modular expert collapse, modular network training instability, modular network overfitting, *etc*. ▷ **(Part 2 — 10 mins) Loop Your CPU and Storage in: Enable Scalable LLM Serving via In-storage Computing.** Moving forward, I will introduce the difficulties of using outdated and budget-constrained hardware to afford LLM serving. To tackle the challenges, I will discuss the possibility of looping CPU and storage into the computation from a full-stack system view. The spotlight here pivots in-storage computing — leveraging a carefully designed multi-level cache management system that could dynamically assign LLM submodules to HBM, DRAM, and SSDs for serving. ▷ **(Part 3 — 10 mins) Against Both Software and Hardware Attacks: Enable Reliable LLM Serving on Risky Outdated GPUs.** As it transitions to serving reliability, I will present the unique safety challenges of serving on risky outdated GPUs beyond classic software perspectives. Specifically, I plan to share our work and recent advances in LLM adversarial serving, jailbreak, hallucination, and hardware-induced stochastic bit-flip errors on outdated GPUs. The goal of my talk is to survey and foster new trends and advances in sustainable and reliable AI, discussing the possibility of recycling outdated GPUs to break the resource monopoly from industries.

References

[1] Chen, T.; Zhang, Z.; JAISWAL, A. K.; Liu, S.; and Wang, Z. 2023. Sparse MoE as the New Dropout: Scaling Dense and Self-Slimmable Transformers. In *The Eleventh International Conference on Learning Representations*.