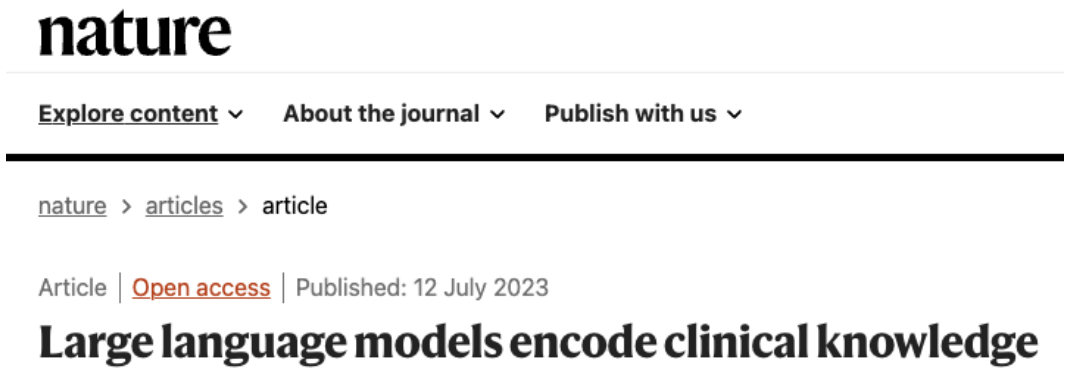
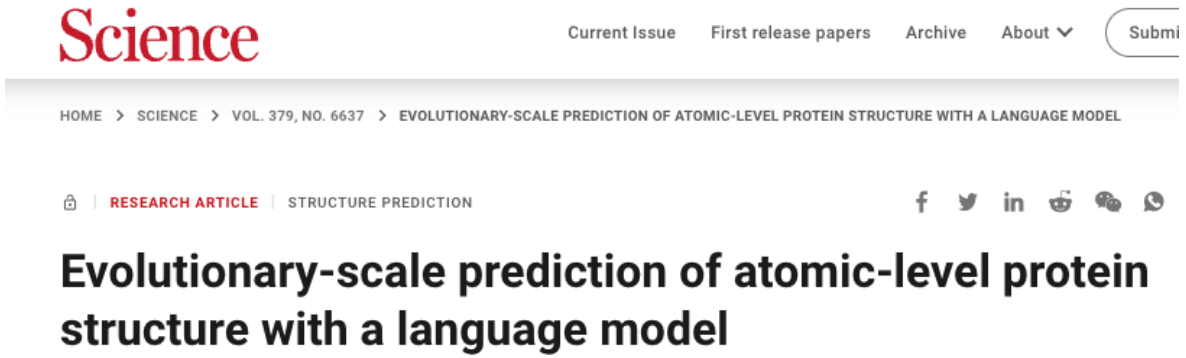
LLM Inference Serving: Survey of Recent Advances and Opportunities

Baolin Li, Yankai Jiang,
Vijay Gadepally, Devesh Tiwari

Surge of LLM-Powered Applications



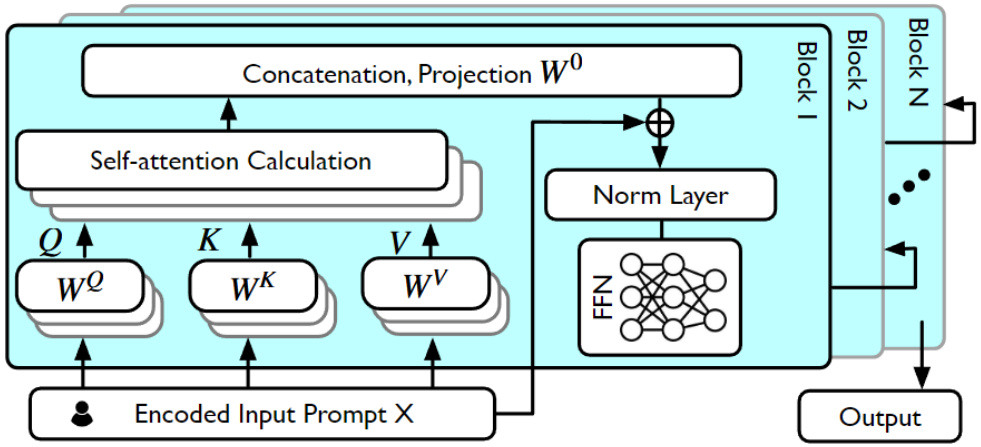


Serving LLMs is Challenging and Rapidly Evolving

We conduct a comprehensive survey of latest LLM serving advances in this paper!

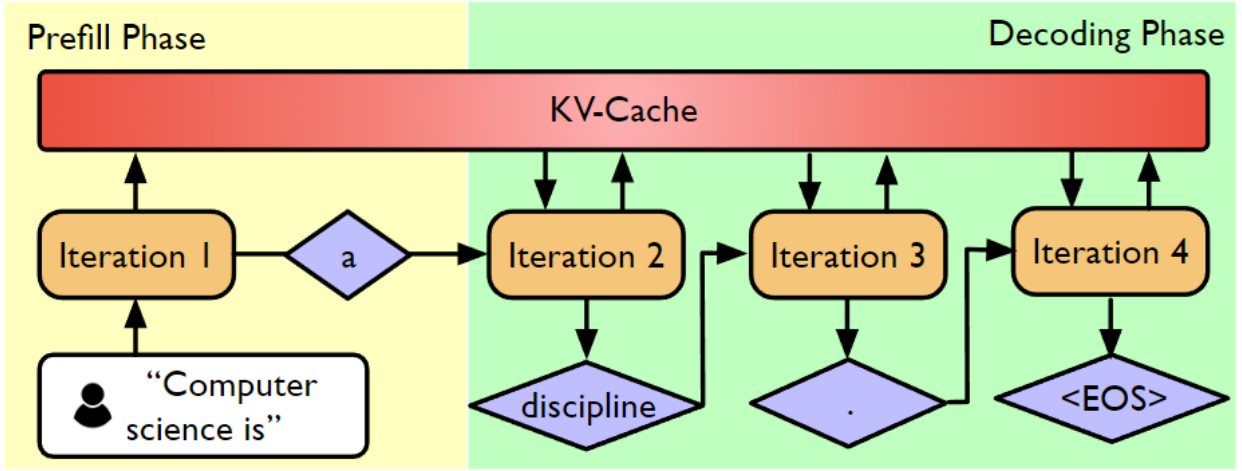
Why LLM Inference?

Transformer architecture



Most popular DNN architecture

Auto-regressive inference



Many new system challenges arise

Overview of this Survey

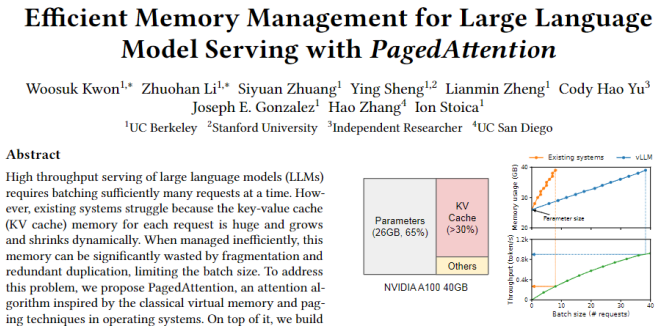
- KV cache and memory management
- Scheduling and computation
- Cloud deployment of LLMs
- Emerging research fields

Many of the research have been integrated into industry-level LLM serving solutions.

We selectively introduce some works in the above areas next.

KV Cache Management

Virtual KV Cache



Efficient Memory Management for Large Language Model Serving with *PagedAttention*

Woosuk Kwon[1,*] Zhuohan Li[1,*] Siyuan Zhuang[1] Ying Sheng[1,2] Lianmin Zheng[1] Cody Hao Yu[3] Joseph E. Gonzalez[1] Hao Zhang[4] Ion Stoica[1]

[1]UC Berkeley [2]Stanford University [3]Independent Researcher [4]UC San Diego

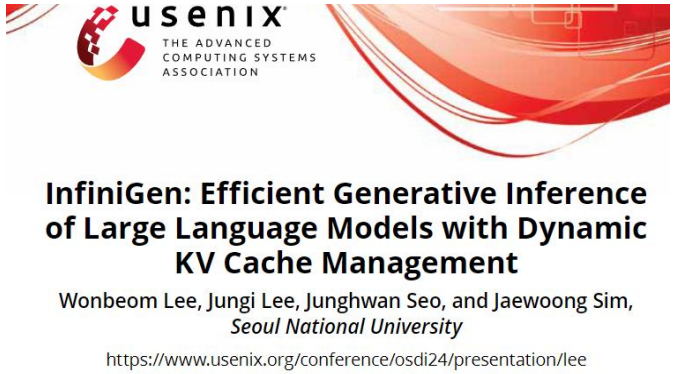
Abstract

High throughput serving of large language models (LLMs) requires batching sufficiently many requests at a time. However, existing systems struggle because the key-value cache (KV cache) memory for each request is huge and grows and shrinks dynamically. When managed inefficiently, this memory can be significantly wasted by fragmentation and redundant duplication, limiting the batch size. To address this problem, we propose PagedAttention, an attention algorithm inspired by the classical virtual memory and paging techniques in operating systems. On top of it, we build

SOSP'23

Manages KV cache in non-contiguous memory blocks that reduces memory fragmentation

Long-Context



usenix THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management

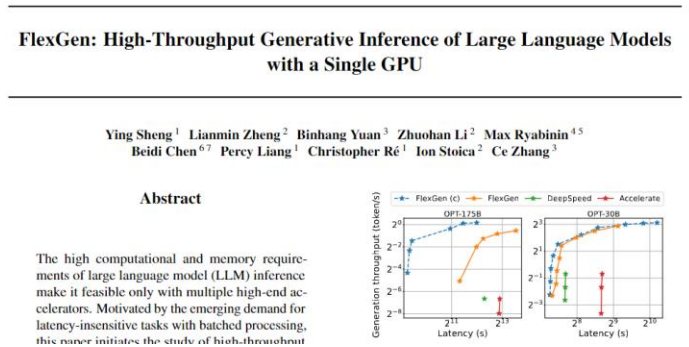
Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim,
Seoul National University

https://www.usenix.org/conference/osdi24/presentation/lee

OSDI'24

Maintains entire KV cache in CPU but only bring a few critical KV caches into GPU to compute attention

KV Cache Compression



FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU

Ying Sheng[1] Lianmin Zheng[2] Binhang Yuan[3] Zhuohan Li[2] Max Ryabinin[4,5] Beidi Chen[6,7] Percy Liang[1] Christopher Ré[1] Ion Stoica[2] Ce Zhang[3]

Abstract

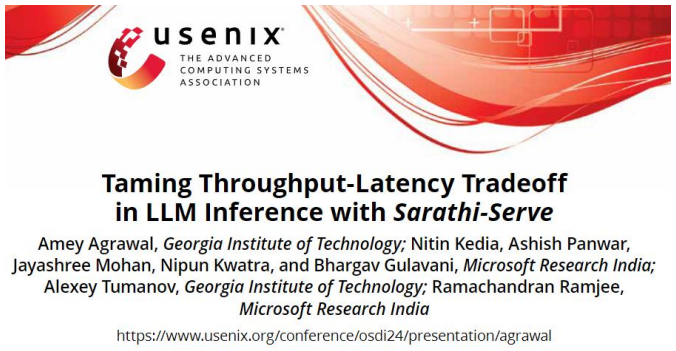
The high computational and memory requirements of large language model (LLM) inference make it feasible only with multiple high-end accelerators. Motivated by the emerging demand for latency-insensitive tasks with batched processing, this paper initiates the study of high-throughput

ICML'23

Aggregates memory and computation from GPU/CPU/disk with 4-bit KV cache compression

Computation Scheduling

Request Batching



OSDI'24

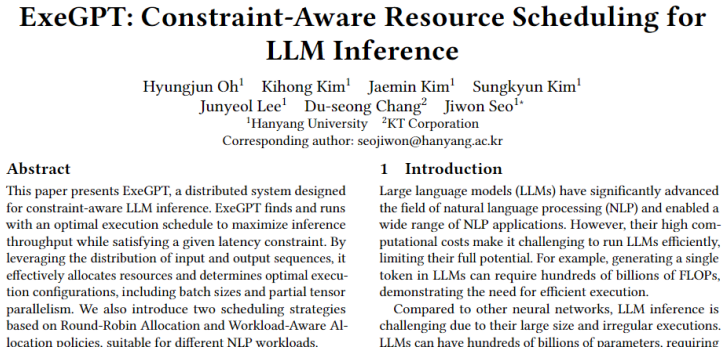
Batching chunked prefill requests with decode requests to achieve better throughput

Disaggregated Inference

OSDI'24

Disaggregate prefill and decode into different inference servers to separately optimize for TTFT and TPOT

Parallelism Optimization


ExeGPT: Constraint-Aware Resource Scheduling for LLM Inference

Hyungjun Oh[1] Kihong Kim[1] Jaemin Kim[1] Sungkyun Kim[1]
Junyeol Lee[1] Du-seong Chang[2] Jiwon Seo[1*]
[1]Hanyang University [2]KT Corporation
Corresponding author: seojiwon@hanyang.ac.kr

Abstract

This paper presents ExeGPT, a distributed system designed for constraint-aware LLM inference. ExeGPT finds and runs with an optimal execution schedule to maximize inference throughput while satisfying a given latency constraint. By leveraging the distribution of input and output sequences, it effectively allocates resources and determines optimal execution configurations, including batch sizes and partial tensor parallelism. We also introduce two scheduling strategies based on Round-Robin Allocation and Workload-Aware Allocation policies, suitable for different NLP workloads.

1 Introduction

Large language models (LLMs) have significantly advanced the field of natural language processing (NLP) and enabled a wide range of NLP applications. However, their high computational costs make it challenging to run LLMs efficiently, limiting their full potential. For example, generating a single token in LLMs can require hundreds of billions of FLOPs, demonstrating the need for efficient execution.

Compared to other neural networks, LLM inference is challenging due to their large size and irregular executions. LLMs can have hundreds of billions of parameters, requiring

ASPLOS'24

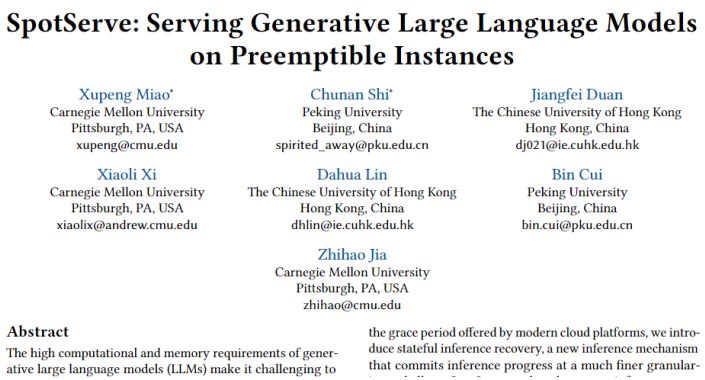
Finds an optimal execution configuration (batch sizes, tensor parallelism) under latency constraint

LLMs in the Cloud

Spot Instance

Serverless

Power Efficiency



SpotServe: Serving Generative Large Language Models on Preemptible Instances

Xupeng Miao*
Carnegie Mellon University
Pittsburgh, PA, USA
xupeng@cmu.edu

Chunan Shi*
Peking University
Beijing, China
spirited_away@pku.edu.cn

Jiangfei Duan
The Chinese University of Hong Kong
Hong Kong, China
dj021@ie.cuhk.edu.hk

Xiaoli Xi
Carnegie Mellon University
Pittsburgh, PA, USA
xiaolix@andrew.cmu.edu

Dahua Lin
The Chinese University of Hong Kong
Hong Kong, China
dhlin@ie.cuhk.edu.hk

Bin Cui
Peking University
Beijing, China
bin.cui@pku.edu.cn

Zhihao Jia
Carnegie Mellon University
Pittsburgh, PA, USA
zhihao@cmu.edu

Abstract

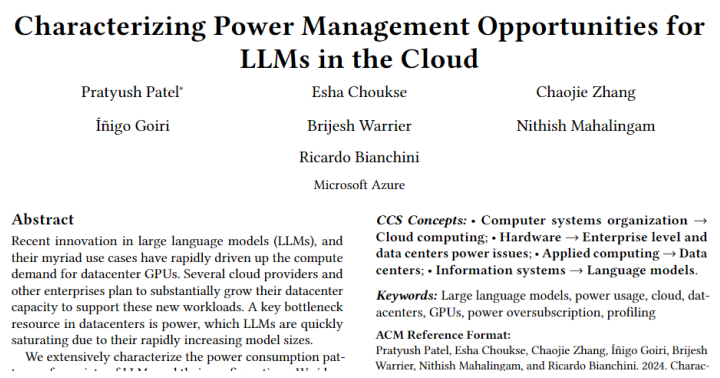
The high computational and memory requirements of generative large language models (LLMs) make it challenging to

the grace period offered by modern cloud platforms, we introduce stateful inference recovery, a new inference mechanism that commits inference progress at a much finer granular-

ServerlessLLM: Low-Latency Serverless Inference for Large Language Models

Yao Fu, Leyang Xue, Yeqi Huang, and Andrei-Octavian Brabete, *University of Edinburgh;* Dmitrii Ustiugov, *NTU Singapore;* Yuvraj Patel and Luo Mai, *University of Edinburgh*

https://www.usenix.org/conference/osdi24/presentation/fu



Characterizing Power Management Opportunities for LLMs in the Cloud

Pratyush Patel*
Esha Choukse
Chaojie Zhang
Íñigo Goiri
Brijesh Warrier
Nithish Mahalingam
Ricardo Bianchini
Microsoft Azure

Abstract

Recent innovation in large language models (LLMs), and their myriad use cases have rapidly driven up the compute demand for datacenter GPUs. Several cloud providers and other enterprises plan to substantially grow their datacenter capacity to support these new workloads. A key bottleneck resource in datacenters is power, which LLMs are quickly saturating due to their rapidly increasing model sizes.

We extensively characterize the power consumption pat-

CCS Concepts: • **Computer systems organization** → **Cloud computing**; • **Hardware** → **Enterprise level and data centers power issues**; • **Applied computing** → **Data centers**; • **Information systems** → **Language models**.

Keywords: Large language models, power usage, cloud, datacenters, GPUs, power oversubscription, profiling

ACM Reference Format:
Pratyush Patel, Esha Choukse, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, and Ricardo Bianchini. 2024. Charac-

ASPLOS'24

OSDI'24

ASPLOS'24

Using spot instance to serve LLMs, minimizes migration cost and resumes at token level

Provides fast inference server loading and locality-aware server allocation strategy to minimize cold start time

Dynamically applies GPU frequency locking and power capping to LLM inference

Emerging Research Areas

RAG

MoE

Sustainability

RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation

Chao Jin[1] Zili Zhang[1] Xuanlin Jiang[1] Fangyue Liu[1]
Xin Liu[2] Xuanzhe Liu[1] Xin Jin[1]
[1]*Peking University* [2]*ByteDance Inc.*

Abstract

Retrieval-Augmented Generation (RAG) has shown significant improvements in various natural language processing tasks by integrating the strengths of large language models (LLMs) and external knowledge databases. However, RAG introduces long sequence generation and leads to high computation and memory costs. We propose RAGCache, a novel multilevel dynamic caching system tailored for RAG. Our analysis benchmarks current RAG systems, pinpointing the performance bottleneck (i.e., long sequence due to knowledge injection) and optimization opportunities (i.e., caching knowledge's intermediate states). Based on these insights, we design RAGCache, which organizes the intermediate states

for generation. With the help of the retrieved documents, RAG expands LLMs' knowledge base and contextual understanding, thereby improving the generation quality [10].

With knowledge injection, RAG introduces long sequence generation for the augmented request, which leads to high computation and memory costs. For instance, the initial request contains 100 tokens, and the retrieved documents may contain 1000 tokens in total. Consequently, the extra computation and memory costs for the augmented request are >10× higher than the original request. This escalation in resource requirements poses a substantial challenge in scaling systems for efficient processing of RAG requests.

Recent work [26, 57], focusing on system optimizations

usenix THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Accelerating Distributed MoE Training and Inference with Lina

Jiamin Li, *City University of Hong Kong;* Yimin Jiang, *ByteDance Inc.;* Yibo Zhu; Cong Wang, *City University of Hong Kong;* Hong Xu, *The Chinese University of Hong Kong*

https://www.usenix.org/conference/atc23/presentation/li-jiamin

Toward Sustainable GenAI using Generation Directives for Carbon-Friendly Large Language Model Inference

Baolin Li*, Yankai Jiang*, Vijay Gadepally†, Devesh Tiwari*
* Northeastern University, † MIT

***Abstract*—The rapid advancement of Generative Artificial Intelligence (GenAI) across diverse sectors raises significant environmental concerns, notably the carbon emissions from their cloud and high performance computing (HPC) infrastructure. This paper presents SPROUT, an innovative framework designed to address these concerns by reducing the carbon footprint of generative Large Language Model (LLM) inference services. SPROUT leverages the innovative concept of "generation directives" to guide the autoregressive generation process, thereby enhancing carbon efficiency. Our proposed method meticulously balances the need for ecological sustainability with the demand for high-quality generation outcomes. Employing a directive**

compute cycles and corresponding carbon footprint. However, it is the inference processes of these LLMs that are poised to become the predominant source of emissions, according to various prior studies [7–9]. Unlike traditional natural language understanding models that predict a single masked word or sentiment, generative LLMs are even more carbon-demanding as they perform iterative predictions for each request until reaching a predefined token or iteration limit. Despite the critical nature of this issue, there's a noticeable gap in research dedicated to reducing carbon emissions specifically

Arxiv'24

ATC'23

EMNLP'24

Cache the intermediate states of the retrieval as a knowledge tree and share across queries

Allocates resources based on expert popularity, optimizes all-to-all communication

Guides the auto-regressive generation to achieve environmental sustainability

Summary and Contributions

This work is the the latest literature survey focusing on LLM inference systems.

We focus solely on system-level performance and efficiency, without alternating the model architecture or retraining.

We hope this survey serves as a valuable resource for LLM system practitioners to stay updated on this rapidly evolving field.

Contact

Baolin Li

li.baol@northeastern.edu

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001, and United States Air Force Research Laboratory Cooperative Agreement Number FA8750-19-2- 1000. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering, or the United States Air Force. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein