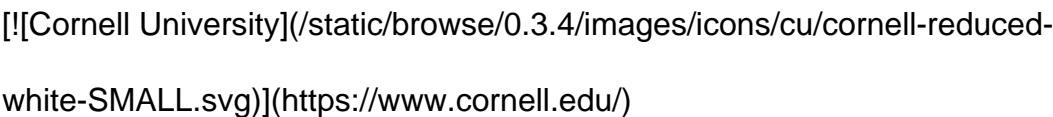


[Skip to main content](#)

 [\(https://www.cornell.edu/\)](https://www.cornell.edu/)

In just 3 minutes help us improve arXiv:

[\[Annual Global Survey\]\(https://cornell.ca1.qualtrics.com/jfe/form/SV_6m22mbqW9GQ3pQO\)](https://cornell.ca1.qualtrics.com/jfe/form/SV_6m22mbqW9GQ3pQO)

We gratefully acknowledge support from the Simons Foundation, [\[member institutions\]\(https://info.arxiv.org/about/ourmembers.html\)](https://info.arxiv.org/about/ourmembers.html), and all contributors. [\[Donate\]\(https://info.arxiv.org/about/donate.html\)](https://info.arxiv.org/about/donate.html)

[\[IgnoreMe\]](#)

 [> \[cs\]\(/list/cs/recent\)](#) > arXiv:2309.06180

[\[Help\]\(https://info.arxiv.org/help\)](https://info.arxiv.org/help) | [\[Advanced Search\]\(https://arxiv.org/search/advanced\)](https://arxiv.org/search/advanced)

All fields Title Author Abstract Comments Journal reference ACM classification
MSC classification Report number arXiv identifier DOI ORCID arXiv author ID
[Help pages](#) [Full text](#)

Search

[![arXiv logo](/static/browse/0.3.4/images/arxiv-logomark-small-white.svg)](https://arxiv.org/)

[![Cornell University Logo](/static/browse/0.3.4/images/icons/cu/cornell-reduced-white-SMALL.svg)](https://www.cornell.edu/)

open search

GO

open navigation menu

quick links

- * [Login](https://arxiv.org/login)
- * [Help Pages](https://info.arxiv.org/help)
- * [About](https://info.arxiv.org/about)

Computer Science > Machine Learning

arXiv:2309.06180 (cs)

[Submitted on 12 Sep 2023]

Title:Efficient Memory Management for Large Language Model Serving with PagedAttention

Authors:[Woosuk

Kwon](<https://arxiv.org/search/cs?searchtype=author&query=Kwon,+W>), [Zhuohan

Li](<https://arxiv.org/search/cs?searchtype=author&query=Li,+Z>), [Siyuan

Zhuang](<https://arxiv.org/search/cs?searchtype=author&query=Zhuang,+S>), [Ying

Sheng](<https://arxiv.org/search/cs?searchtype=author&query=Sheng,+Y>), [Lianmin

Zheng](<https://arxiv.org/search/cs?searchtype=author&query=Zheng,+L>), [Cody

Hao Yu](<https://arxiv.org/search/cs?searchtype=author&query=Yu,+C+H>), [Joseph

E.

Gonzalez](<https://arxiv.org/search/cs?searchtype=author&query=Gonzalez,+J+E>),

[Hao Zhang](<https://arxiv.org/search/cs?searchtype=author&query=Zhang,+H>),

[Ion Stoica](<https://arxiv.org/search/cs?searchtype=author&query=Stoica,+I>)

View a PDF of the paper titled Efficient Memory Management for Large Language

Model Serving with PagedAttention, by Woosuk Kwon and 8 other authors

[View PDF](/pdf/2309.06180)

- > 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
- > vLLM, an LLM serving system that achieves (1) near-zero waste in KV cache
- > memory and (2) flexible sharing of KV cache within and across requests to

> further reduce memory usage. Our evaluations show that vLLM improves the
> throughput of popular LLMs by 2-4 \times with the same level of latency
> compared to the state-of-the-art systems, such as FasterTransformer and
> Orca. The improvement is more pronounced with longer sequences, larger
> models, and more complex decoding algorithms. vLLM's source code is publicly
> available at [this https URL](https://github.com/vllm-project/vllm)

Comments: | SOSP 2023

---|---

Subjects: | Machine Learning (cs.LG); Distributed, Parallel, and Cluster Computing (cs.DC)

Cite as: | [arXiv:2309.06180](https://arxiv.org/abs/2309.06180) [cs.LG]

| (or [arXiv:2309.06180v1](https://arxiv.org/abs/2309.06180v1) [cs.LG] for this version)

| <<https://doi.org/10.48550/arXiv.2309.06180>> Focus to learn more arXiv-issued DOI via DataCite

Submission history

From: Woosuk Kwon [[view email]](/show-email/2fbc22fc/2309.06180)]

[v1] Tue, 12 Sep 2023 12:50:04 UTC (831 KB)


Full-text links:

Access Paper:

View a PDF of the paper titled Efficient Memory Management for Large Language

Model Serving with PagedAttention, by Woosuk Kwon and 8 other authors

- * [\[View PDF\]\(/pdf/2309.06180\)](/pdf/2309.06180)
- * [\[TeX Source\]\(/src/2309.06180\)](/src/2309.06180)
- * [\[Other Formats\]\(/format/2309.06180\)](/format/2309.06180)

[ (<https://arxiv.org/icons/licenses/by-4.0.png>) view license
(<http://creativecommons.org/licenses/by/4.0/> "Rights to this article")

Current browse context:

cs.LG

[< prev](/prevnext?id=2309.06180&function=prev&context=cs.LG "previous in cs.LG \(\accesskey p\)\") | [next >](/prevnext?id=2309.06180&function=next&context=cs.LG "next in cs.LG \(\accesskey n\)\")

[new](/list/cs.LG/new) | [recent](/list/cs.LG/recent) | [2023-09](/list/cs.LG/2023-09)

Change to browse by:

[cs](/abs/2309.06180?context=cs)
[cs.DC](/abs/2309.06180?context=cs.DC)

References & Citations

- * [\[NASA ADS\]\(https://ui.adsabs.harvard.edu/abs/arXiv:2309.06180\)](https://ui.adsabs.harvard.edu/abs/arXiv:2309.06180)
- * [\[Google Scholar\]\(https://scholar.google.com/scholar_lookup?arxiv_id=2309.06180\)](https://scholar.google.com/scholar_lookup?arxiv_id=2309.06180)
- * [\[Semantic Scholar\]\(https://api.semanticscholar.org/arXiv:2309.06180\)](https://api.semanticscholar.org/arXiv:2309.06180)

[1 blog link](/tb/2309.06180)

([what is this?](https://info.arxiv.org/help/trackback.html))

[a](/static/browse/0.3.4/css/cite.css) export BibTeX citation Loading...

BibTeX formatted citation

×

loading...

Data provided by:

Bookmark

[![BibSonomy logo](/static/browse/0.3.4/images/icons/social/bibsonomy.png)

](http://www.bibsonomy.org/BibtexHandler?requTask=upload&url=https://arxiv.org/abs/2309.06180&description=Efficient

Memory Management for Large Language Model Serving with PagedAttention

"Bookmark on BibSonomy") [![Reddit

logo](/static/browse/0.3.4/images/icons/social/reddit.png)

](https://reddit.com/submit?url=https://arxiv.org/abs/2309.06180&title=Efficient

Memory Management for Large Language Model Serving with PagedAttention

"Bookmark on Reddit")

Bibliographic Tools

Bibliographic and Citation Tools

Bibliographic Explorer Toggle

Bibliographic Explorer [_\(\[What is the Explorer?\]\(https://info.arxiv.org/labs/showcase.html#arxiv-bibliographic-explorer\)\)_](https://info.arxiv.org/labs/showcase.html#arxiv-bibliographic-explorer)

Connected Papers Toggle

Connected Papers [_\(\[What is Connected Papers?\]\(https://www.connectedpapers.com/about\)\)_](https://www.connectedpapers.com/about)

Litmaps Toggle

Litmaps [_\(\[What is Litmaps?\]\(https://www.litmaps.co/\)\)_](https://www.litmaps.co/)

scite.ai Toggle

scite Smart Citations [_\(\[What are Smart Citations?\]\(https://www.scite.ai/\)\)_](https://www.scite.ai/)

Code, Data, Media

Code, Data and Media Associated with this Article

alphaXiv Toggle

alphaXiv [_\(\[What is alphaXiv?\]\(https://alphaxiv.org/\)\)_](https://alphaxiv.org/)

Links to Code Toggle

CatalyzeX Code Finder for Papers [_\(\[What is CatalyzeX?\]\(https://www.catalyzex.com\)\)_](https://www.catalyzex.com/)

DagsHub Toggle

DagsHub [_\(\[What is DagsHub?\]\(https://dagshub.com/\)\)_](https://dagshub.com/)

GotitPub Toggle

Gotit.pub [_\(\[What is GotitPub?\]\(http://gotit.pub/faq\)\)_](http://gotit.pub/faq)

Huggingface Toggle

Hugging Face [_\(\[What is Huggingface?\]\(https://huggingface.co/huggingface\)\)_](https://huggingface.co/huggingface)

Links to Code Toggle

Papers with Code [_\(\[What is Papers with Code?\]\(https://paperswithcode.com/\)\)_](https://paperswithcode.com/)

ScienceCast Toggle

ScienceCast _([What is ScienceCast?](https://sciencecast.org/welcome))_

Demos

Demos

Replicate Toggle

Replicate _([What is Replicate?](https://replicate.com/docs/arxiv/about))_

Spaces Toggle

Hugging Face Spaces _([What is
Spaces?](https://huggingface.co/docs/hub/spaces))_

Spaces Toggle

TXYZ.AI _([What is TXYZ.AI?](https://txyz.ai))_

Related Papers

Recommenders and Search Tools

Link to Influence Flower

Influence Flower _([What are Influence
Flowers?](https://influencemap.cmlab.dev/))_

Core recommender toggle

CORE Recommender [_\(\[What is CORE?\]\(https://core.ac.uk/services/recommender\)\)_](https://core.ac.uk/services/recommender)

IArxiv recommender toggle

IArxiv Recommender [_\(\[What is IArxiv?\]\(https://iarxiv.org/about\)\)_](https://iarxiv.org/about)

- * Author
- * Venue
- * Institution
- * Topic

About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [\[**Learn](#)

more about arXivLabs**](https://info.arxiv.org/labs/index.html).

[Which authors of this paper are endorsers?](/auth/show-endorsers/2309.06180) | [Disable MathJax](javascript:setMathjaxCookie\(\)) ([What is MathJax?](https://info.arxiv.org/help/mathjax.html))

* [About](https://info.arxiv.org/about)

* [Help](https://info.arxiv.org/help)

* contact arXivClick here to contact arXiv [Contact](https://info.arxiv.org/help/contact.html)

* subscribe to arXiv mailingsClick here to subscribe [Subscribe](https://info.arxiv.org/help/subscribe)

* [Copyright](https://info.arxiv.org/help/license/index.html)

* [Privacy Policy](https://info.arxiv.org/help/policies/privacy_policy.html)

* [Web Accessibility Assistance](https://info.arxiv.org/help/web_accessibility.html)

* [arXiv Operational Status](https://status.arxiv.org)

Get status notifications via

[email](https://subscribe.sorryapp.com/24846f03/email/new) or

[slack](https://subscribe.sorryapp.com/24846f03/slack/new)