

2025 Fall Systems Reading Group

Welcome Everyone!

Zhihui Chen, Ouxiang Zhou and Ruibo Liu

2025.09.16

Agenda

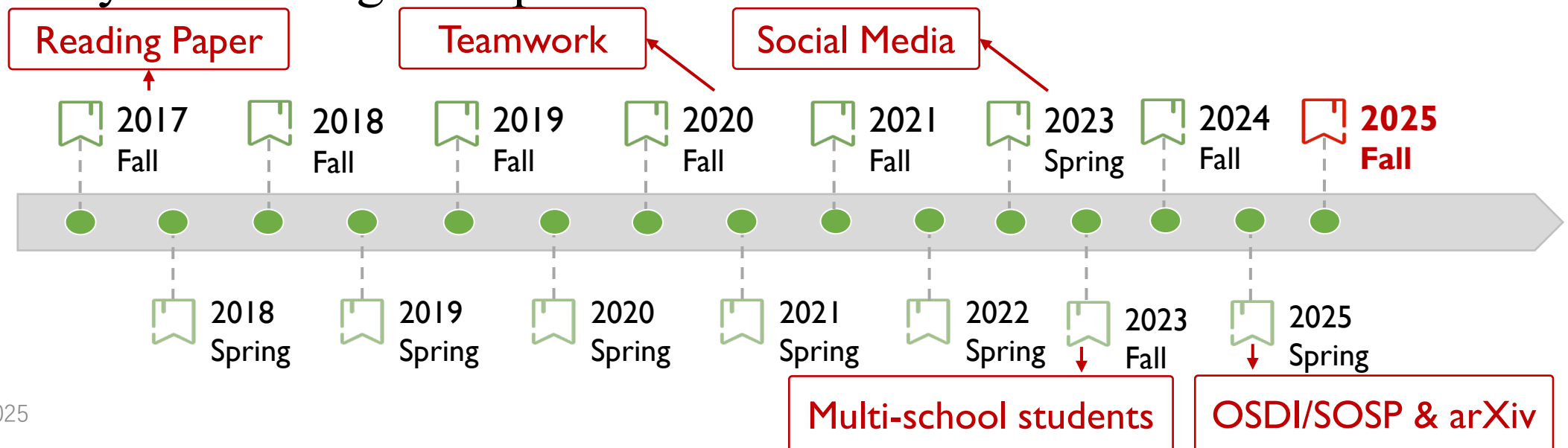
- **Introduction to Reading Group**
 - Mission
 - Arrangement
 - Format & Requirements
- Advices for reading a paper
- Advices for giving a talk

Mission of reading group

- Understand and keep abreast of “latest research in **systems research**”
- Learn “how to do **high-quality** systems research”
- Polish soft skills
 - Understanding
 - Presentation
 - Writing
 - Critical thinking
 - Communication
 - ...

Mission of reading group

- Understand and keep abreast of “latest research in **systems research**”
- Learn “how to do **high-quality** systems research”
- History of Reading Group



Mission of reading group

- Understand and keep abreast of “latest research in **systems research**”
- Learn “how to do **high-quality** systems research”
- Target of this semester
 - **Paper Sharing**
 - Improve the presentation quality
 - More discussion and brainstorming
 - Improve writing skills
 - **More than one paper**
 - Choose one more paper from arXiv

Previous RG

- We read papers from:
 - SOSP' 24
 - OSDI' 24, 25
 - arXiv
- 16 presentations were given
- Presenters were from
 - USTC ADSL
 - UESTC
 - ...

Schedule

February 25

- 🕒 Kick-off meeting
- 👤 Jiyang Wang, Kunzhao Xu and Cheng Li
- 📄 slides

March 11

- 🕒 Comprehensive introduction of DeepSeek-AI's technical report (PART I)
- 👤 Xin Ren, Tonghuan Xiao, Jiahui Tan, Yandong Shi, Kunzhao Xu, Yifei Liu, Chongzhao Yang, Jiaan Zhu, Zewen Jin, Yinhe Chen, Ping Gong, Guanbin Xu, Haiquan Wang, Quan Zhou and Chaoyi Ruan
- 📄 MLA slides, 📄 DualPipe slides, 📄 FP8 Training slides, 📄 MTP slides
- 🗂 Q&A summary, 🎥 video

March 18

Topic I

- 🕒 Comprehensive introduction of DeepSeek-AI's technical report (PART II)
- 👤 Xin Ren, Tonghuan Xiao, Jiahui Tan, Yandong Shi, Kunzhao Xu, Yifei Liu, Chongzhao Yang, Jiaan Zhu, Zewen Jin, Yinhe Chen, Ping Gong, Guanbin Xu, Haiquan Wang, Quan Zhou and Chaoyi Ruan
- 📄 RL slides, 📄 3fs slides
- 🗂 Q&A summary, 🎥 video

Topic II

- 🕒 [OSDI'24] Ladder: Enabling Efficient Low-Precision Deep Learning Computing through Hardware-aware Tensor Transformation
- 👤 Chengru Yang
- 📄 slides
- 🗂 Q&A summary, 🎥 video

March 25

Topic I

- 🕒 [OSDI'24] FairyWren: A Sustainable Cache for Emerging Write-Read-Erase Flash Interfaces
- 👤 Qingyuan Chen
- 📄 slides

Topic II

2025 Spring

Specific Requirements

Other Information

Schedule

February 25

March 11

March 18

Topic I

Topic II

March 25

Topic I

Topic II

Summary and Video

April 1

Topic I

Topic II

Summary and Video

April 8

Topic I

Topic II

April 15

April 22

Topic I

Topic II

Summary and Video

April 29

Topic I

Topic II

Summary and Video

May 6

Topic I

Topic II

May 13

Topic I

Topic II

Summary and Video

May 20

Summary and Video

Previous RG

- Topic
 - Storage / Memory
 - Vector search
 - Tiered memory
 - Disaggregated memory
 - File system
 - Cloud computing
 - LLM / AI
 - RAG
 - Scheduling
 - KV Cache
 - Parallelism
 - Low-precision computation
 - DeepSeek-AI's technical report
 -

ADSL Reading Group

Schedule

February 25

- 🕒 Kick-off meeting
- 👤 Jiyang Wang, Kunzhao Xu and Cheng Li
- 📄 slides

March 11

- 🕒 Comprehensive introduction of DeepSeek-AI's technical report (PART I)
- 👤 Xin Ren, Tonghuan Xiao, Jiahui Tan, Yandong Shi, Kunzhao Xu, Yifei Liu, Chongzhao Yang, Jiaan Zhu, Zewen Jin, Yinhe Chen, Ping Gong, Guanbin Xu, Haiquan Wang, Quan Zhou and Chaoyi Ruan
- 📄 MLA slides, 📄 DualPipe slides, 📄 FP8 Training slides, 📄 MTP slides
- 📄 Q&A summary, 📺 video

March 18

Topic I

- 🕒 Comprehensive introduction of DeepSeek-AI's technical report (PART II)
- 👤 Xin Ren, Tonghuan Xiao, Jiahui Tan, Yandong Shi, Kunzhao Xu, Yifei Liu, Chongzhao Yang, Jiaan Zhu, Zewen Jin, Yinhe Chen, Ping Gong, Guanbin Xu, Haiquan Wang, Quan Zhou and Chaoyi Ruan
- 📄 RL slides, 📄 3fs slides
- 📄 Q&A summary, 📺 video

Topic II

- 🕒 [OSDI'24] Ladder: Enabling Efficient Low-Precision Deep Learning Computing through Hardware-aware Tensor Transformation
- 👤 Chengru Yang
- 📄 slides
- 📄 Q&A summary, 📺 video

March 25

Topic I

- 🕒 [OSDI'24] FairyWren: A Sustainable Cache for Emerging Write-Read-Erase Flash Interfaces
- 👤 Qingyuan Chen
- 📄 slides

Topic II

2025 Spring

Specific Requirements

Other Information

Schedule

February 25

March 11

March 18

Topic I

Topic II

March 25

Topic I

Topic II

Summary and Video

April 1

Topic I

Topic II

Summary and Video

April 8

Topic I

Topic II

April 15

April 22

Topic I

Topic II

Summary and Video

April 29

Topic I

Topic II

Summary and Video

May 6

Topic I

Topic II

May 13

Topic I


Topic II

Summary and Video

May 20

Summary and Video

What do we read?



19th USENIX Symposium on Operating Systems Design and Implementation

JULY 7-9, 2025
BOSTON, MA, USA

Co-located with **USENIX ATC '25**

Sponsored by USENIX in cooperation with ACM SIGOPS

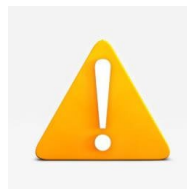


SOSP 2025
The 31st Symposium on Operating Systems Principles

October 13 – 16, 2025 · **Lotte Hotel World**, Seoul, Republic of Korea

Early Registration deadline: September 8, 2025 (previously September 1, 2025)

Online Registration: [\[Cvent Link\]](#)



Read best papers!!!

What do we read?



Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention

Jingyang Yuan^{*1,2}, Huazuo Gao¹,
Y. X. Wei¹, Lean Wang¹, Zhiping

²Key Laboratory for Multimed

{yuanjy, mzhang_cs}@pk

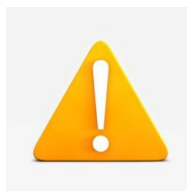
KIMI K2: OPEN AGENTIC INTELLIGENCE

TECHNICAL REPORT OF KIMI K2

Kimi Team

ABSTRACT

We introduce Kimi K2, a Mixture-of-Experts (MoE) large language model with 32 billion activated parameters and 1 trillion total parameters. We propose the MuonClip optimizer, which improves upon Muon with a novel QK-clip technique to address training instability while enjoying the advanced token efficiency of Muon. Based on MuonClip, K2 was pre-trained on 15.5 trillion tokens with zero



Read latest papers!!!

Paper sharing: arrangement

- Time: 19:00 – 21:00, every Tuesday
- Location:
 - Offline: 高新区信智楼A707
 - Online: Tencent meeting 877-6724-4752
- Webpage: https://adsl-rg.github.io/2025_fall.html

Paper sharing: arrangement

- Time: 19:00 – 21:00, every Tuesday

- Location:

- Offline: 高新区信智
- Online: Tencent meet

- Webpage: <https://adsl-rg.g>

2025 Fall


Specific Requirements

- We focus on the latest papers from SOSP and OSDI, as well as papers released on arXiv. Each time presenters select one paper from SOSP or OSDI and one from arXiv.
- The presentation follows a "1+N" format, where one person delivers the main content while supporting members assist with preparation and manage the Q&A session. These supporting members are also encouraged to contribute to the presentation.
- The discussion should provide a thorough analysis of the paper's strengths and weaknesses, along with a comprehensive review of related work from the past three years. The presentation must be at least 45 minutes long.

Other Information

The playback video and text summary will be uploaded to [bilibili](#) and [zhihu](#) as soon as possible.

Paper sharing: arrangement

- Each presentation led by two students
 - Choose the paper (one paper from OSDI or SOSP and one from arXiv)
 - Find your teammates (one team for OSDI/SOSP paper and the other for arXiv)
 - **Guarantee the quality**
 - Presentation video: Upload to 
- We also encourage students from other schools or labs to participate in the RG :)

Paper sharing: format

- Primary focus: **understanding the paper**
 - What is the problem?
 - What are state-of-the-arts, and their deficiencies?
 - What are the challenges?
 - What are the key insights/techniques?
 - Lessons learned from experiments?
- Whole discussion: 1.5~2 hours, presentation: **70~80 minutes**

Paper sharing: tips

- Please make around **70 slides!**
 - Too much text ☹
 - Copy paste figures ☹
 - Animations ☺
 - Transitions between slides ☺
- One slide: 1 - 2 minutes
- Please do rehearsals offline

Paper sharing: tips

- Please make around **70 slides!**
 - Too much text ☹️
 - Copy paste figures ☹️
 - Animations 😊
 - Transitions between slides 😊
- One slide: 1 - 2 minutes
- Please do rehearsals offline
- Additional requirement:
 - **A mind map**
 - **Summary after sharing**
 - Problem
 - Key insights/techniques
 - Evaluation
 - Strengths
 - Improvement
 - Record Q&A (by Ouxiang & Ruibo)
 - Submit to 知 (by Ouxiang & Ruibo)

Ready to share?

- Please make around **70 slides!**
 - Too much text ☹️
 - Copy paste figures ☹️
 - Animations 😊
 - Transitions between slides 😊
- Additional requirement:
 - **A mind map**
 - **Summary after sharing**
 - Problem
 - Key insights/techniques
 - Evaluation

Ready to share? Fill the **follow document!**

<https://docs.qq.com/sheet/DRGVKV3NEcHJGTnpz?tab=BB08J2>

If you are from other schools or labs, let us know :)

Agenda

- Introduction to Reading Group
 - Mission
 - Arrangement
 - Format & Requirements
- **Advices for reading a paper**
- Advices for giving a talk

How to read a paper!

- From Srinivasan Keshav
 - The Robert Sansom Professor of Computer Science at the University of Cambridge
 - ACM/IEEE Fellow
- **Three passes**
 - 1st: get a bird's-eye view
 - 2nd: grasp the content
 - 3rd: rethink, recreate the work
- <http://ccr.sigcomm.org/online/files/p83-keshavA.pdf>



Agenda

- Introduction to Reading Group
 - Mission
 - Arrangement
 - Format & Requirements
- Advices for reading a paper
- **Advices for giving a talk**

Advices

- <https://people.eecs.berkeley.edu/~jrs/speaking.html>
 - Preparing a talk
 - Giving the talk
- <http://pages.cs.wisc.edu/~markhill/conference-talk.html>
 - Oral presentation advice
 - How to give a bad talk

2025 Fall Systems Reading Group

Q&A